DOCUMENT RESUME

ED 217 724	FL 013 014

AUTHOR	Culhane, Terry, Ed.; And Others
TITLE	Practice and Problems in Language Testing. Proceedings of the International Language Testing Symposium of the Interuniversitare Sprachtestgruppe (4th, Essex, England, September 14-17, 1981).
INSTITUTION	Essex Univ., Colchester (England). Dept. of Language and Linguistics.
PUB DATE	Sep 81
NOTE	169p.; In its: Occasional Papers No. 26. Not available in paper copy because of poor reproducibility of original document. One article in German has been omitted. For related documents, see FL 013 015-028.
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Cloze Procedure; Communicative Competence (Languages); Computer Assisted Testing; English (Second Language); Error Analysis (Language); Higher Education; Language Attitudes; *Language Tests; *Predictive Validity; Rhetoric; *Second Language Learning; Spanish; *Standardized Tests; Teacher Attitudes; *Testing; *Test Validity; Writing Evaluation

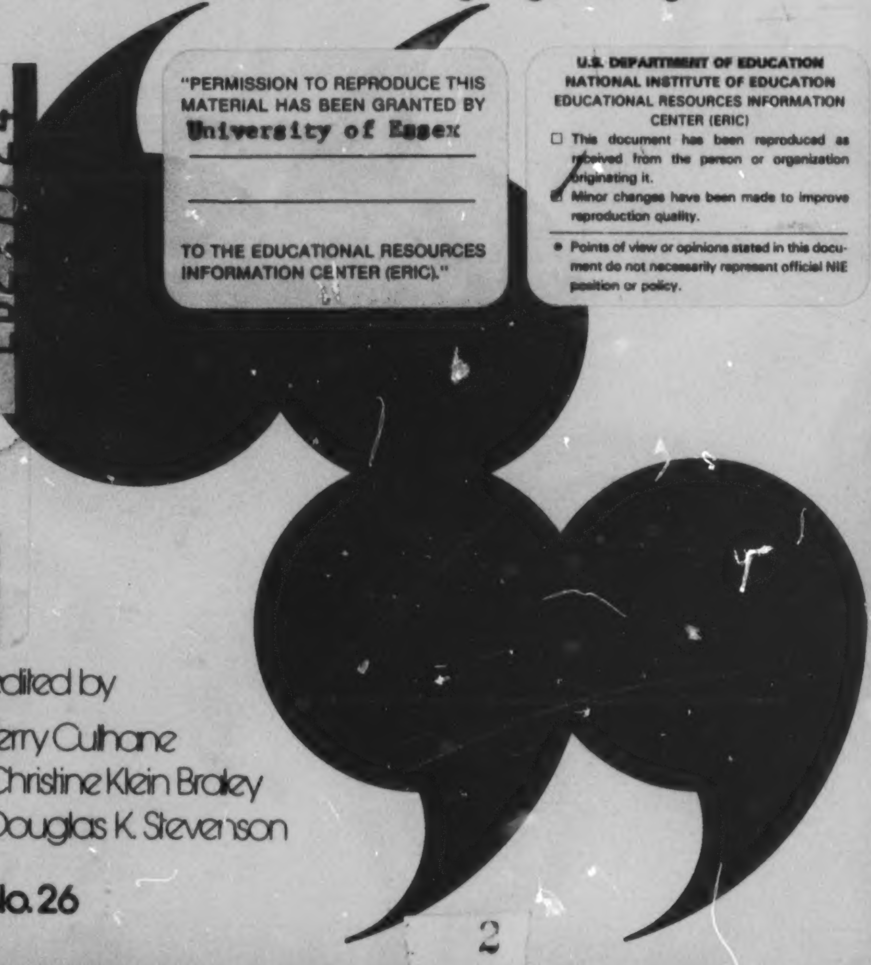
ABSTRACT

These proceedings on language testing include the following papers: (1) "The Style Test," by H. Bonheim; (2) "Testing between the Cultures: A Case Study of ESP Testing in Non-idealized Situations," by R. Brunt; (3) "Graded Objectives and Tests in British Schools," by M. Buckby; (4) "The Spanish Attributive Construction SER/ESTAR + Adjective: Real Use Phenomena Submitted to Adult Learners," by N. Delbecque and J. De Kock; (5) "Towards a Practical Model of a Communicative Competence Testing Program," by R. Fischer; (6) "Report on an Experiment in Activated Group Achievement Testing for Advanced Students of Spoken English," by T. Hayward, A. Michiels, and J. Noel; (7) "The Defense Language Institute Foreign Language Center Achievement Test Program," by R. Henderson; (8) "Computer-Based Analysis of Open-Ended Foreign-Language Test Items," by T. Lindblad; (9) "Testing Oral Proficiency in Life-Like Settings: Rating Scales and Reliability," by J. Mets; (10) "The Test of Ability to Subordinate: Predictive and Concurrent Validity for Attained ESL Composition," by K. Perkins; (11) "The C-Test--A Modification of the Cloze Procedure," by U. Raatz and C. Klein-Braley; (12) "A Review of Issues Raised in the Production of the English Language Testing Service," by I. Seaton; (13) "Teachers' Attitudes towards Language Tests and Testing," by D. Stevenson and U. Riewe; and (14) "A Study of the Predictive Validity of the Michigan Battery, with Special Reference to the Composition Component," by D. Wall. (AMH)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

University of Essex
Department of
Language and Linguistics

Occasional Papers
Practice and problems in language testing



ED217724

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
University of Essex

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.
☑ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

FL013014

edited by
Terry Culhane
Christine Klein Braley
Douglas K. Stevenson

No. 26

Practice and Problems in Language Testing

Practice and Problems in Language Testing

Proceedings of the Fourth International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe, held at the University of Essex, 14 - 17th September, 1981.

Editors:
Terry Culhane (Essex)
Christine Klein-Braley (Duisburg)
Douglas K. Stevenson (Essen)

Foreword

The Symposium started with a workshop on testing, the purpose of which was to provide background information for the uninitiated. Most of the people who registered for this section were, however, far from innocent beginners in the field. Nevertheless, the opportunity to re-articulate basic ideas on reliability, validity, test construction and basic statistics proved a stimulating and useful process for those who attended this workshop, and an effective introduction to the main Symposium, the proceedings of which are contained in this volume.

We have arranged the papers in alphabetical order, since to have done otherwise would have required a good reason, and we were unable to think of one.

Our thanks are due to the authors of the papers for adhering so meticulously (some more so than others!) to the style sheet, and thus making our task an easier one; to Kate Hampson, Graduate Secretary of the Department of Language and Linguistics, University of Essex, for helping with the organisation of the Symposium; to Barbara Hopkins, who gave so much help with the manuscripts. Finally, we wish particularly to thank Ulrike Riewe, IUS member from Essen, for her help with the volume.

Preface

The Fourth International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe was the first of the various "regional" and "international" meetings to be held outside of continental Europe. The fact that participants came from as far away as California, the Philippines, and Thailand indicates that the IUS enjoys support which also goes beyond its established base in Europe. More importantly, the strongly positive remarks we have received about this meeting indicate that the language testing needs that exist, internationally, are well-worth trying to meet, despite the difficulties involved in organizing such a symposium.

The IUS membership, therefore, wishes to express its sincere gratitude to Terry Culhane of the University of Essex, who served as Chairman for the Symposium, to the British Council, which generously gave of its own expertise and also kindly provided financial assistance, and not least, to the administration, staff, and personnel of the University who made our stay in Colchester such a pleasant one.

Christine Klein-Braley
Douglas K. Stevenson

Co-Chairmen, IUS

Contents

The Style Test 1
Helmut Bonheim

Testing between the Cultures: A Case Study of ESP Testing in Non-idealized Situations 10
Richard J. Brunt

Graded Objectives and Tests in British Schools 18
Michael Buckby

The Spanish Attributive Construction SER/ESTAR + Adjective : Real Use Phenomena Submitted to Adult Learners 24
N. Delbecque and J. De Kock

Towards a Practical Model of a Communicative Competence Testing Program 35
Robert A. Fischer

Report on an Experiment in Activated Group Achievement Testing for Advanced Students of Spoken English 48
T. Hayward, A. Michiels, and J. Noel

The Defense Language Institute Foreign Language Center Achievement Test Program 58
Robert G. Henderson

Computer-based Analysis of Open-ended Foreign-Language Test Items 75
Torsten Lindblad

Pages 88-94, which contained an article in German, have been omitted.

Testing Oral Proficiency in Life-like Sattings: Rating Scales and Reliability 95
Jan Mets

The Test of Ability to Subordinate: Predictive and Concurrent Validity for Attained ESL Composition 104

Kyle Perkins

The C-Test - A Modification of the Cloze Procedure 113

Ulrich Raatz and Christine Klein-Braley

A Review of Issues Raised in the Production of the English Language Testing Service 139

Ian Seaton

Teachers' Attitudes towards Language Tests and Testing 146

Douglas K. Stevenson and Ulrike Riewe

A Study of the Predictive Validity of the Michigan Battery, with Special Reference to the Composition Component 156

Dianne Wall

ED217725

THE STYLE TEST

Helmut Bonheim (University of Cologne)

Items on English style have been used since 1975 in eight different versions of the Cologne placement test for beginning university students. And a separate style section has been included in six successive versions of a test battery designed for the German Academic Exchange Service. This test is administered annually to between 400 and 500 students who have applied for a stipend to support a period of study in Great Britain. The style sections have been more than satisfactory in both kinds of test.

At the outset we assumed that grammar, vocabulary and reading comprehension formed the core of a language test, and that style was a comparatively marginal area. We discovered, however, that this assumption is questionable. Apparently the style test can play a distinctive role in a foreign language test battery.

We began with an item format borrowed from the vocabulary subtest. The head consists of a sentence or two with a blank in it, and this blank is to be filled with one of four options. The option may be a single word, as in a vocabulary test, or a phrase, clause or sentence. The style item also differs from a vocabulary item in that all the options are semantically correct. But only one of them fits the style of the context. In other words, it tests not grammaticality but acceptability and suitability to a context:

> My husband and I are writing you to thank you for inviting us to the delightful supper and ______ concert last Friday night.
>
> A) ripping
> B) splendid
> C) great
> D) high-grade

This format has the advantage of simplicity and efficiency. But in practice we found the "sore finger" item preferable:

> My parents were asleep at last, $\underset{A}{\underline{\text{thank goodness}}}$. I could hear my father $\underset{B}{\underline{\text{snoring his head off}}}$ in the next room. $\underset{C}{\underline{\text{I sneaked downstairs}}}$, $\underset{D}{\underline{\text{proceeding forth}}}$ into the garden, and ran across

FL013015

the lawn to where my lover was waiting.

Items in this format are easier to compose. For the item with a blank, four words or expressions have to be found with more or less the same meaning, but only one of them is to fit the context adequately. The sore finger format, on the other hand, requires that we insert only a single unacceptable element into a text, either an authentic one or one made up for the purpose. So it is not difficult to write satisfactory items. The sore finger item just quoted turned out to be a little too easy for students in their third or fourth semesters. It had a facility index of .805 (n= 472). But the discrimination index reached .40 and the use of the distractors was as balanced as could be expected for such an easy item. It was therefore transferred to the test for beginning students. Here (PT Form N, administered October, 1978) the facility index dropped about 20 points and the discrimination index rose to .56 in one administration (n= 547) and .614 in another (n=524). The reliability of the style subtest consisting of 25 questions reached .88 (KR20) and showed a mean discrimination index of .51. In fact, of the 14 style items included in our very first venture into the field in 1975, 9 of the 14 items had a discrimination index of above .4 (n=595), 4 of them over .6. These are results we hardly ever achieve with our vocabulary or idiom items.

The sore finger format has a cousin in the odd man out format. Here a set of texts is presented, one of which is of a different author, date or genre from the others:

Which lines were written in the eighteenth century?

A) Row after row with strict impunity
The headstones yield their names to the element,
The wind whirrs without recollection.

B) God moves in a mysterious way,
  His wonders to perform;
He plants his footsteps in the sea,
  And rides upon the storm.

C) O mortal folk! you may behold and see
  How I lie here, sometime a mighty knight;
The end of joy and all prosperity
  Is death at last, thorough his course and might.

D) Where are Elmer, Herman, Bert, Tom and Charley,
The weak of will, the strong of arm, the clown,
  the boozer, the fighter?
All, all, are sleeping on the hill.

The correct choice is B: the opening lines of William Cowper's "Light Shining out of Darkness." It is to be chosen because of its contrast to the clearly twentieth century texts by Allen Tate (A) and Edgar Lee Masters (D) on the one hand and the Tudor epitaph by Stephen Hawes (C) on the other.

This format makes sense. It has the advantage for students of literature that a variety of clues can be operative--metrical, semantic, and so forth--so that the everyday situation of a reader is to some extent simulated. But the format also has disadvantages. For one thing, it requires a number of separate passages of greater length than do the other two formats. For another, the distractors must derive from and be seen by the candidate to derive from a distinctly different source or set of sources than does the key (Pynsent 1970). The odd man out is therefore both a little cumbersome and rather difficult to write.

Although both the format with a blank and the odd man out texts have turned out to be useful alternatives, the sore finger format is the most satisfactory in practice; it is used in 150 of the 253 style items which are in our item pool at the moment.

The style item, as we see, can be cast in formats not normally used in objective tests. Nor are the mental operations those called for in other subtests:

--vocabulary, in contrast to style, is something our pupils often work up systematically in the course of their school training in English--style is certainly not;

--grammar can also be systematically studied, and

--reading comprehension, unlike style, demands a sense of logical connections which we work hard to get our students to master but over which we seem to have little influence.

Style, by contrast, is a domain almost impossible to master systematically, and logic seems to be of little use in solving a style item. The contrast to reading comprehension is suggested by the fact that candidates require only about a third of the time for a style item that they do for a reading comprehension item (Bolz 1978). Evidently the right answer, as in a vocabulary item, is either immediately evident or it is not: thinking the problem over for a while does not help.

A number of other findings suggest the special status of the style test:

1 Retest results

Scores in reading comprehension seem to be remarkably stable on re-testing. But scores on the style subtest can change quite dramatically. In the spring of 1976 we used a set of sub-tests with a group of 425 students in their third semester, sub-tests we had used a few months earlier in a test of 592 university entrants. The third semester students were 15% better than the beginners in reading comprehension but 78% better in the sub-test entitled style and verbal logic (Bonheim 1976). Since the two tests used and the populations as well were not strictly comparable, the matter was investigated again with a set of re-tests involving students in Cologne and in Bonn (Jackson 1979). In each of three groups of candidates in the first, second and third year of study the improvement in the style score was quite striking. Indeed, in the group of candidates who had spent five or more weeks in an English speaking country after first taking the test, the improvement in style was twice that of the improvement in reading comprehension. Although the number of candidates in this group was relatively small (n=16), the results were significant at the 1% level. Somewhat different results were arrived at in a study by Gellert and Schindler (1978), who used our test to re-examine 72 students who had spent half a year in Britain. The students were retested about a year after they had returned to Germany, in other words, two whole years after the first administration of the test. These students, a selected group of scholarship holders who had spent several months at a British university, gained significantly in all areas of the test. Although the gain in style was greater than in grammar and three other areas, the gain in vocabulary and reading comprehension was even greater yet (Gellert and Schindler 1978). The overall gains were significant at the .5% level.

Subtest	1976	1978	gain
Vocabulary	70.7%	88.5%	17.8%
Idiom	59.0	70.5	12.5
Style	66.9	81.6	14.7
Literary Terms	77.8	89.4	11.6
Literature	57.8	78.7	20.9
Reading Comprehension	53.7	77.0	23.3
Grammar	67.7	79.5	11.8
British Life	67.5	81.5	13.5

As the table suggests, the gain in style is still greater than in two areas which are generally taught in our German university language courses, grammar and literary terminology. We arrive at the paradoxical conclusion that our students' sense of style, which is not developed through any systematic training, nevertheless tends to improve quite appreciably.

2 Factor analysis

Our first test including a style section was analyzed using the procedure of the SPSS suite of computer programmes. Three distinct factors were identified. These three factors encompassed 45% of the total variance of the test, with factor number 1, namely the style factor, accounting for 43.3% of the common variance (Massey 1976). This was a surprising result. It suggests that the role of style in foreign language competence is not so marginal as we had initially supposed.

3 Prognostic validity

The style section of our entrance test had a somewhat higher prognostic validity than did the other subtest, using as a measure of success the intermediate examination taken at the end of the second year of study (Bonheim and Kreifelts 1979).

4 Local variations

We have found that the range of results in the style test is remarkably different from university to university (Kreifelts 1976). This is also suggested by the high standard deviation of scores. Whereas the standard deviation of the whole of the test in question amounted to about a quarter of the total score, the standard deviation of the style sub-test was more than half as high as the mean score itself:

results of Placement Test H1, 1975 (n = 592)

	Total	Grammar	Reading Compr.	Vocab.	Style
Score	64.5	19.6	11.7	22.7	10.5
S. D.	16.9	4.0	3.6	7.7	6.2

Most of the scores in other subtests achieved by candidates in the ten universities where this test was used varied only a little from the average, but the results on the style section showed quite extraordinary variations. This suggests what had been suspected, that the intake of students at some institutions is markedly different from that at others. It may be that the marked differences in preparation for university work level out later: we do not know. The hypothesis may be allowed that such levelling out is unlikely.

5 Style and general achievement

This hypothesis is perhaps supported by the fact that the results of our style test showed a correlation of .412 with the verbal factor of a general achievement test (Hornsches Leistungsprüfsystem), whereas the correlation of this test with the grammar section was an insignificant .115 and with the reading comprehension section .289. Only the vocabulary and the style sections correlated significantly with marks in that section of the general achievement test which is likely to remain fairly stable in a young adult (Bonheim and Kreifelts 1979). Thus although some of our evidence suggests that students' sense of style improves remarkably compared to other language skills, other evidence links the results of our style test with mental factors possessing a degree of stability. Perhaps there is a solution to the paradox: the verbal factor is relatively stable in the native language, but necessarily in a much more fluid state in a foreign language learner. If our thesis that competence in style is an important factor in foreign language skills is valid, the question whether it is stable or amenable to influence is worth further consideration. But at the moment we know very little about how style competence develops in the native language, and still less about its growth in relation to a foreign language.

Style in tests and in teaching

Having said all this, we have not yet said what style is, especially in reference to what aspects of it can be caught in an objective test item. Style always involves a genuine choice of grammatically permissible options, whereas grammar does not; for the distractors in a grammar item are ungrammatical, the distractors in a style item may be grammatically and lexically acceptable, but unsuitable. So the grammar, vocabulary and reading comprehension items all ask: "What is correct?" The style item asks: "What is suitable?" In other words, in the framework of a foreign language test style is a question of offering more or less correct options, only one of which is likely to be chosen by a native speaker or writer in that particular context. So style turns out to be a matter of taste, of probabilistic rather than of generalisable rules of occurrence, and to contain some of the sociolinguistic components which Spolsky thinks appropriate to the "post-modern" age of language testing.

Of course language which is acceptable in the lecture hall may not be acceptable in the pub. So the tester must create a context by which the candidate can judge the options on offer. And the testing situation demands that the context be a brief text, the briefer the better. It follows that the text must display a maximum degree of stylistic homogeneity. That homogeneity can be established in various ways. The most common basis is that of style register: many of our items consist of a poetic text into which an unpoetic expression has been inserted:

In a field by the river my love and I did stand,
And on my leaning shoulder she laid her $\underset{A}{\underline{\text{snow-white}}}$
    hand.

She $\underset{B}{\underline{\text{bid}}}$ me take life easy, as the grass $\underset{C}{\underline{\text{luxuriates}}}$
    on the weirs;
But I was young and $\underset{D}{\underline{\text{foolish}}}$, and now am full of
    tears.

The fact that "luxuriates" fails to fit the metrical pattern may offer an additional clue, but it is surprising how often such clues are ignored.

About a quarter of our items are based on this poetic/unpoetic contrast. The relationship can of course be turned on its head: we insert into a colloquial or slangy text a literary or stilted element:

> Dearest Dolly,
>
> You aren't, I do hope, angry with me for bringing you home a little late last night. What a delightful party! But your dear old pater $\underset{A}{\underline{\text{glared}}}$ out the upstairs window while I was backing Jittery Jemina into the street again, so I $\underset{B}{\underline{\text{fancy}}}$ he was none too pleased about my $\underset{C}{\underline{\text{dropping}}}$ you at the doorstep a little later than I pledged to do.
>
> Yours $\underset{D}{\underline{\text{faithfully}}}$,
>
> Jack

Here the candidate is expected to know that a letter to "Dearest Dolly" will not end with "Yours faithfully".

A number of other bases for stylistic homogeneity can be chosen by the item writer. He can mismatch refined with vulgar, standard with dialect, spoken with written, archaic with contemporary, academic with journalese, pop song cliché with bureaucratic jargon, and so forth. Perhaps a well thought-out specification grid ought to be constructed, on the basis of which items can be commissioned and later assembled into subtests. But in every case the challenge presented must be a matter not of what is right or wrong in the narrower linguistic sense but what would be likely or acceptable in a particular context.

Style deserves a more careful consideration in both school and university teaching of English. For our present teaching materials attack only *some* of the elements in the standard taxonomies of educational objectives, namely analysis and synthesis. Evalution is rarely called for. Perhaps there is a certain justification for this neglect: after all, a feeling for style, like other kinds of evaluation, is learnable, but hardly teachable, and apparently grows quite measurably even in students who are not consciously taught in this domain. Nevertheless the question arises: if the sense of style improves by itself, what rewards might we reap if nature were given a little help?

If we wanted to put together a course in remedial stylistics, we might design some of its materials in the light of how our style items work. For instance, we could work with so-called "text-mockeries." These are texts in which, as in a sore finger item, a few key expressions have been replaced by unsuitable locutions-- unsuitable because they stem from another century, another author, the language of a different social class, etc. The text mockery is an exercise akin to a sore finger test item, except that one takes a longer text and the student must find the odd men out on his own. These are not marked like the options in a multiple choice item. (The cloze test is similar, except that there one deletes items at intervals rather than replacing them.) The text mockery is a useful therapeutic measure which is easy to prepare. It can be based on the texts relevant to the course or seminar, and in practice it proves to have great entertainment value both for the teacher and for the student. Unlike an objective style item, it does not take great expertise to prepare. Its effectiveness can be investigated by means of an objective style test, in which items focussed on the same or on similar problems have been included.

The domain of the style test may well interest the linguist as well as the teacher. Many of our items, for instance, call attention to the social context of utterances, which grammar tests rarely do. What we mean by communicative competence surely includes a sense of the social context, of the situational rather than merely linguistic acceptability of an utterance, of the aesthetic effect of one expression as opposed to another, of the stiltedness of Smith's conversation as opposed to the vivid slanginess of Jones'. The mastery of vocabulary and grammar is not enough to confer near-native Englishness on our students. Admittedly, grammatical competence is still a central and essential part of communicative competence, but stylistic competence is something else in addition: an integrative skill with a performance aspect. Perhaps attention to matters of style is in fact essential if we want to help our students achieve a near-native performance in a foreign language.

Bibliography

Bolz, N. 1978. Vorwissen und Problemlösen: zwei bedingende Variablen für die Bearbeitungszeit von Testitems. Cologne: Arbeitsgruppe Sprachtests, Bericht Nr. 24.

Bonheim, H. 1978. Ein Universitätseingangstest für Neuphilologen. Cologne: Arbeitsgruppe Sprachtests, Bericht Nr. 18.

Bonheim, H. and B. Kreifelts. 1979. Ein Universitätseingangstest für Neuphilologen. Bonn.

Canale, M. and M. Swain. 1980 "Theoretical Basis of Communicative Approaches to Second Language Teaching and Testing." Applied Linguistics 1: 1-47.

Gellert C. and G. Schindler. 1978. Erfolgsanalyse zum Anglistenprogramm. Munich: Bayrisches Staatsinstitut für Hochschulforschung und Hochschulplanung.

Hymes, D. 1972. "On communicative competence". In Pride, J. B. and Holmes, J. (eds). Sociolinguistics. Harmondsworth.

Jackson, S. 1979. "The Development of Foreign Language Competence among German Students of English." Cologne: Arbeitsgruppe Sprachtests. Bericht 26; also in: Beiträge zur Hochschulforschung 4: 351-363.

Kreifelts, B. 1976. Auswertung der Tests F2 und H1. Cologne: Arbeitsgruppe Sprachtests, Bericht 17.

Massey, A. 1976. An Initial Study of the Factorial Validity of Placement Test H1. Cologne: Arbeitsgruppe Sprachtests, Bericht 19: repr. in Bonheim and Kreifelts 1979.

Pynsent, R. 1970. Objective Style/Register Tests. Cologne: Arbeitsgruppe Sprachtests, Bericht 9.

ED217726

TESTING BETWEEN THE CULTURES: A CASE STUDY OF ESP TESTING IN NON-IDEALIZED SITUATIONS

Richard J. Brunt (University of Essen)

1 Introduction

Before a test can be designed and administered certain essential questions have to be asked and answered. Among these are the following:

why test?
to what use will the test be put?
for what population is the test intended?
what content based upon what syllabus is the test to have?
what information is to be gained from the test and how is this going to be used?
what practical constraints will apply?

It must be assumed that all these questions can be asked and satisfactorily answered, otherwise there is little point in going on to define test format and procedure. Should the teacher and/or test designer receive satisfactory answers, he must also assume that both the students taking the particular course for which the testing is intended and the staff teaching it are convinced of its relevance and worth, and therefore share a positive motivation. Further, the educational objectives of such a course must be clearly defined and definable. There must, in short, be a good fit between the aims of the course and the chosen test.

In the case of EFL teaching with a general or literary orientation at an institute of tertiary education, one should be able to make the following assumptions: that students are learning English to fulfil some specific need, that there is a syllabus or text book upon which the course is based, that what students are learning can be tested, that a test can be constructed to do this, that the test results will, for example, be used to diagnose learning problems or to assess achievement, and that the test will be demonstrated to have both reliability and validity.

If we move from the area of "general" to "specific" English, then these assumptions can no longer be made with any certainty. In this paper, I am concerned with that branch of English for Specific Purposes (ESP) usually referred to as English for Science and Technology (EST). This area has undergone a rapid expansion in the last ten years. Up to the late sixties, science students, if they learned English at all, normally attended courses which taught English through literary texts and so were better acquainted with Wordsworth than with the English of welding or woodwork. Strevens (1971 a) was among the first to point out the illogicality of this state of affairs and to call for an 'alternative to daffodils'.

FL013016

ESP acknowledges that teaching should be related to the needs of the students and to the subject being studied. The logicality of this position has now been generally accepted, although some would maintain there is no such thing as 'specific' English, and there is still much debate on a more difficult question, namely whether EST is merely 'general' English with the addition of some specialized vocabulary, or whether it is a separate variety with its own rules and conventions (cf. Corbluth 1975; Blackie 1975).

Initially, it was hence not only unclear *what* English science students should study, but also how much they needed and what skills they should acquire or possess. Although much research has been done in the meantime to clarify the nature of ESP (for a summary, cf. Beier 1980) and to define the needs of science students, there is still a great degree of uncertainty about what we are doing when we teach a physicist English, whether in his own country or in England or America. It was even suggested earlier this year at a SELMOUS conference held at the University of Essex that many of us are engaged in TENOR - Teaching English for No Obvious Reason! It may be that the physicist does not *need* English or, put another way, "the more physics one knows, the less English one needs" (Spolsky 1973: 166,fn 8). On the other hand, there is a common, if often unspoken, belief that "Scientists are [...] philistines, or at the very least 'uncivilised', and any activity that smacks of measurement or quantification is low-valued" (Strevens 1971 a:9). There are those who consider uneducated anyone who doesn't know Nabokov, Updike or Doris Lessing, yet themselves have never read, for example, Eddington, Asimov or Freeman Dyson. The chasm of Snow's 'Two Cultures' yawns unbridgably, with the omniscent magus with his incomprehensible ideas on the one side, and the barbarian, unkissed by the civilizing graces of the humanities on the other. This view is so extreme as to be ridiculous, but does in a more moderate version still influence attitudes towards ESP teaching.

Although exceptions can be found, it is still not common for language teachers and testers to be trained in the pure or applied sciences. They still predominantly receive a literary training which reinforces the belief that scientific texts are somehow inferior to the works of the great novelists, or at least deficient in 'style'. This lack of appreciation of the scientific way of thinking and the unawareness of the nature of scientific English frequently make it difficult, if not impossible, for a teacher or tester of ESP, especially if not trained in the area, to even know what questions to ask. As one member of the project described by Selinker (1979:201) exclaimed in frustration: "'How can we teach the NNS [non-native speaker] to read this stuff when we don't even know what we don't even know!'" As Lackstrom, Selinker and Trimble have shown in many studies, the teaching of EST, especially reading, does not merely involve the provision of new lexis and grammar, but rather demands a knowledge of the rhetoric and presuppositions of written scientific discourse, which may or may not be present in the first language competence. And, as Dubois (1981) has shown, spoken scientific discourse also has its own rhetorical and structural conventions, for example, the need to persuade an audience of one's credibility. These differ from those of every day speech, and are as yet largely unexplored. There is, in fact, still much to be discovered in the field of ESP/EST, but I wish now to turn from the future to the present and to consider ESP teaching and testing in various contexts.

2 ESP teaching - assumptions for testing

ESP/EST is taught both as a Second Language and as a Foreign Language. In the first case, it is possible that EST is taught from the start, that is, science is learnt in English. Alternatively, ESP teaching starts after several years of general English, and this may be in the students' own country or in an English-speaking country. In EFL situations there is a great degree of variation, but most European countries appear to require science students to learn some English - and perhaps Russian.

What and how much English is taught obviously depends upon what priorities are set. They range from only reading comprehension to reading and writing skills, to only spoken English, for example, laboratory, workshop or service English (cf. Munby 1978), each approach being in response to different needs. The basic question is, however, whether science students need English at all. It is normally asserted that, since English is the lingua franca of all the sciences, especially in the case of journals, which often accept articles only in English, irrespective of the author's nationality, a science student needs to be able to read English, and later on to be able to write it as well. It is also maintained that spoken English is necessary, particularly in ESL situations, for communication in the laboratory or workshop and later either to sell one's products or to discuss one's contracts in English-speaking countries, or to enable one to read a paper and to chat to one's colleagues at international conferences.

All these reasons are in theory sufficient to justify the teaching of the English specific to a scientific or technological discipline, but I am doubtful whether they all apply to the specific situation at one German university, at which EST teaching is required. I wish now to show the problems which arise within an institutional framework when needs, aims and methods have not been sufficiently clarified. I do not suggest that my remarks are representative of the situation at universities as a whole. On the other hand, one thing that is generalizable is that in real language teaching and testing situations, the specific problems are the most common.

3 Case Study

In the *Bundesland* of North-Rhine Westphalia there are five recently-founded universities called *Gesamthochschulen*, or comprehensive universities. They aim to offer the growing number of students requiring higher education the opportunity to study the arts and sciences under one roof. Unlike the older universities, they are open to students who do not have the general *Abitur*, the German equivalent of A-levels. They are able to study economics, mathematics, physics, chemistry, mechanical and civil engineering, and industrial and communication design with the *Fachabitur* gained at a technical college. Since they concentrate here on practical work in technical subjects, there is less concentration on English, German and mathematics, and this fact led the educational authorities to stipulate that they must attend additional courses in these subjects.

These are the so-called *Brückenkurse* or "bridging courses", intended to compensate for the deficits in non-grammar school students' knowledge of 'basic subjects'. Two studies, one by Hitpaß of the University of Cologne and the other by a group at the University of Essen, suggest that any deficits, if present, do not seem to correspond with the type of school previously attended. Furthermore, it appears that non-grammar school pupils tend to produce better results in the applied sciences than their grammar-school contemporaries. Nevertheless, these courses were established by legal decree in 1978, and in the case of English the following stipulations were made:

- there must be 100 hours of classwork, including practice periods
- the course must be completed before the students are allowed to sit their intermediate examination
- the course should take place as a "compact course" before the beginning of the term
- the course must be *studiengangbezogen*, that is, relevant to the subject of study
- a *Klausur* or written examination lasting four hours is required to complete the course
- in the case of failure, the examination may be repeated once
- the "test" must involve the comprehension of a 500-word English text which has not been previously discussed in class
- students have passed the final examination if they have achieved a "satisfactory performance" (*ausreichende Leistung*)

In addition, the conditions for the contents of the courses and the final examinations are determined by the Ministries of Science and Education, which may also demand to inspect the questions in the final examination and the marked scripts.

The "comprehensive universities" are legally obliged to implement these stipulations, but in practice there have been some attempts with the support of the university administrations, to make them more in accord with reality, especially as regards the problem of fitting them into a very full timetable. For instance, at the University of Essen the courses take place during the semester rather than before it, and students are allowed to take them up to their fourth semester. In response to the protests of students and staff, some changes are being considered by the Ministries of Science and Education, namely:

- 80 hours per course
- the additional optional subjects history, French, physics and chemistry, on top of compulsory mathematics or English
- the planning of the courses and the final examination by the universities themselves.

At the moment, however, it is still the case that a university cannot decide if there will be a test, and, if so, what form it will take. It has been decided that the only possible form is a four-hour written examination with five questions on the content of the text to be answered in English. The educational authorities have decreed that only this "test" will produce the required results.

Before going on to consider the question of "institutional validity" (Stevenson 1981), it is worth noting what is not in the ministerial decree. EST is not mentioned, but, since all three *Brückenkurse* are required to be relevant to the subject of study, this is assumed to mean in the case of English that specific English, for example, English for chemists, is to be taught rather than conversational English. How this is to be taught is not stated - there is no explicit syllabus, nor are textbooks specified. The only aim towards which teaching could be directed is the final test. This, however, has for all intents and purposes been decreed valid. There is no evidence that connects the test to the goals of the class, or either of these to the needs of the students. We are, therefore, still in the position of putting a blind horse before a cart without wheels. Since no pre-testing can be required, no selection of students is possible. One is faced, in theory, with up to 350 students a year, all of whom must take English, irrespective of their previous knowledge or present abilities. Until recently, African students, for example, were obliged to take the courses, although they had had all of their secondary education in English, just like Korean students who had little German and less English.

The educational authorities, then, have decided how much English students need and what will be regarded as a valid test, yet have only stated that the degree of proficiency must be "sufficient". It is not surprising, therefore, if many students feel that the Ministery is not primarily interested in what they have learned and for what purpose, but in having an "ultimate deterrent" which will ensure that some students think twice about studying for eight semesters. As the teacher of these courses I am obliged to observe the ministerial stipulations. At the same time I must, of course, gain as much latitude as possible so that courses can be better matched to the needs of my students.

The students themselves are thoroughly in favour of learning the English of their subject. They admit they will have to read English books and articles in the course of their studies. In practice, however, the *Brückenkurse*, irrespective of their content, are, as I have indicated, regarded by many of the students obliged to take them as a political manoeuvre on the part of the educational authorities - a deliberate barrier to prevent them from taking the degree of their choice, despite the evidence that they are no less able than their grammar-school colleagues. In the past three years, students have actively opposed the courses through organised protests both at the university and regional level. The courses have been attended by less than half of those obliged to take them. Some of those who have regularly participated have nevertheless boycotted the final *Klausur*. Those who do take it are aware of its arbitrary nature and are only interested in receiving the *Schein*, the certificate which will allow them to sit for their intermediary examination.

As regards the actual teaching and the question of the appropriate content for a suitable test, I have found that few of the standard ESP/EST text books are adequate to the needs of these students. For example, in many books the reading comprehension texts are too elementary or grammar-based. In many cases, the students' subject-knowledge in German is far in advance of their proficiency in English. Their needs can be more closely met by using texts from school text books of an 'O' to 'A' level standard or of American high school standard, and by supplementing them with grammar and vocabulary exercises. A rough analysis of the final written examinations has revealed certain errors which occur with a high degree of frequency and which can be made the basis of teaching in future courses. Otherwise, such a *Klausur* has proved of little value in assessing students' reading comprehension. This is hardly surprising, given the fact that the test was prescribed independent of any goals or needs analysis.

4 Conclusion

As I have tried to show, some of the essential questions which have to be asked about testing are pre-empted in the situation at Essen. "Why test?" "Because the law says so". "What test?" "The one we tell you to use". "What information for what purpose?" "We don't know, so you don't need to know". What is valid and useful has been decided a priori. I don't wish to give the impression, however, that I have no freedom or am merely the agent of the educational authorities. Rather, it is my task and challenge to make the best of the given conditions and to determine what the students need to learn. Yet the variety in their backgrounds and needs makes the teaching and testing task extremely difficult.

The students have normally had three to six years of English at secondary level, but normally have not come into contact with EST. This does not present a problem as regards specialized vocabulary, since words such as *electrolysis*, *isomeric*, *molecule* or *stoichiometric* are either present in German in a slightly different orthography or can be easily learned. Rather, and my experience here coincides with that of Cowan (1974) in Iran and Cohen et al. (1979) in Israel, it is the lack of knowledge of the sub-technical vocabulary containing such 'hard words' as *deduce*, *infer*, *suppose*, *increase*, *occur*, etc., which inhibits comprehension. Yet it is precisely these words which carry the main burden of scientific argumentation. Testing implies some specification of what these words are. To this end, I have compiled a list of some 600 items, using such well-known frequency lists as Thorndike/Lorge, West and Kučera/Francis (Brown Corpus). Since the corpora from which they are derived represent various stylistic registers, and may underrepresent the scientific register (cf. Cowan 1974:391), I have also drawn on frequency lists of scientific English such as Zettersten (1969) and Hoffmann (1970). In addition, the glossaries of ESP text books and a limited corpus based on the texts used by the students were included. In this way, it is hoped that participants are given a solid basis which enables them to then acquire the specialized vocabulary of their subject. The content validity of this list rests, of course, on expert judgement.

As regards syntax and morphology, it is undoubtedly true that some constructions occur statistically more frequently than others in scientific texts, notably the passive (cf. Gerbert 1970, Beier 1980:77-80). But it is unproductive, I think, to concentrate on teaching this and other forms such as modal verbs, conditional clauses and participal constructions without showing why and when they occur, and what argumentative and rhetorical functions they perform in the text as a whole. Students must be made aware of the fact that, for example, article usage differs between scientific and non-scientific contexts (cf. Lackstrom, Selinker and Trimble 1970:110-112), and that such words as *therefore, however, thus* and *finally* are not meaningless fillers, but are often used with a different semantic and rhetorical function, namely to mark the structure of the argument as 'grammatico-logical operators' (Strevens 1971 b) or 'syntactic markers of cohesion' (Halliday and Hasan 1976; cf. Cohen et al. 1979: 558-560).

The cloze, interestingly, has proved of some value in testing students' awareness of the logical cohesion between sentences. A somewhat illegitimate relative of the cloze, that is the blanking of only such cohesion markers, has also given useful results. Similarly, reading tests, both in multiple-choice form and as transfer exercises, for example, text to graph or table, have been tentatively used to determine where reading problems lie.

There is still some way to go before the questions posed by Scholz and Scholz "What evaluative instruments should be used? What subject matter should the test contain (general EFL or ESP?) What do those instruments measure - science or language? Which instrument provides a reasonable indicator of a learner's future performance in science and technology?" (Scholz and Scholz 1979:207) can be categorically and exhaustively answered for EFL/EST learners in general, and for those in Germany in particular. It is to be hoped, however, that in the near future, now that the planning of the teaching and testing is gradually being relinquished by the Ministries and becoming the responsibility of the universities, it will no longer be a case of "If you are in this course, then you must need English", but rather "I am studying EST because I have the following clearly defined needs which can best be met by a course with the following objectives. The test helps me and the teacher to assess my learning difficulties and to predict my success".

Bibliography

Beier, R. 1980. Englische Fachsprache. Stuttgart etc.: Kohlhammer

Blackie, D.J.J.S. 1979. "Towards a definition of ESP". English Language Teaching Journal 33/4: 262-266.

Cohen, A. et al. 1979. "Reading English for Specialized Purposes: Discourse analysis and the use of student informants". TESOL Quarterly 13/4: 551-564.

Corbluth, J.D. 1975. "English? - or 'special English'?". English Language Teaching 29: 277-286.

Cowan, J.R. 1974. "Lexicon and syntactic research for the design of EFL reading materials". TESOL Quarterly 8/4: 389-399.

Dubois, B.L. 1981. "Nontechnical arguments in biomedical speeches". Perspectives in Biology and Medicine 24/3: 399-410.

Gerbert, M. 1970. Besonderheiten der Syntax in der technischen Fachsprache des Englischen. Halle: VEB Max Niemeyer Verlag.

Halliday, M.A.K. and R. Hasan. 1976. Cohesion in English. London: Longman

Hoffmann, L. 1970. Fachwortschatz Physik: Häufigkeitswörterbuch, Russisch, Englisch, Französisch. Leipzig: VEB Verlag Enzyklopädie.

Lackstrom, J.E., Selinker, L. and L.P. Trimble. 1970. "Grammar and technical English". In Lugton, R.C. (ed.). English as a second language: Current issues. Philadelphia, PA: Center for Curriculum Development. 101-133.

Munby, J. 1978. Communicative syllabus design. Cambridge: Cambridge University Press.

Scholz, G.E. and C.M. Scholz. 1979. "Testing in an EFL/ESP context". In Yorio, C.A., Perkins, K. and J. Schachter (eds.). On TESOL '79. The learner in focus. Washington DC: TESOL. 206-229.

Selinker, L. 1979. "On the use of informants in discourse analysis and 'Language for Specialized Purposes'.". International Review of Applied Linguistics 17/3: 189-215.

Spolsky, B. 1973. "What does it mean to know a language; or how do you get someone to perform his competence?". In Oller, J.W. and J.C. Richards (eds.). Focus on the learner: pragmatic perspectives for the language teacher. Rowley, MA: Newbury House. 164-176.

Stevenson, D.K. 1981. "Beyond faith and face validity: the multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests". In Palmer, A.S., Groot, P.J.M. and G.A. Trosper (eds.). The validation of oral proficiency tests: An introduction. Washington, DC: TESOL.

Strevens, P. 1971 a. "Alternatives to daffodils. Or scientist thou never wert". In Perren, G.E. (ed.). Science and technology in a second language. London: CILT. 7-11.

Strevens, P. 1971 b. "The medium of instruction (mother tongue/second language) and the formation of scientific concepts". International Review of Applied Linguistics 9/3: 267-274.

Zettersten, A. 1969. A word-frequency list of scientific English. Lund: Studentlitteratur.

ED217727

GRADED OBJECTIVES AND TESTS IN BRITISH SCHOOLS

Michael Buckby (University of York)

At the last count, there were over 60 groups throughout the U.K. developing schemes of graded objectives and tests. There was a clear need to pool experiences - which has been undertaken by CILT - and to evaluate the effects of these schemes on pupils, parents and teachers - which has been the task of a research project at the Language Teaching Centre at York, funded by the Schools Council. The full report of this research was published in 1981 and is available from CILT: "Graded objectives and tests for modern languages: an evaluation." The report is presented in a way designed to help and encourage discussions between foreign language teachers and others interested in "the graded test movement". The results, which are summarised below, are extremely positive and encouraging, while presenting important challenges to foreign language teachers both in the U.K. and elsewhere.

At the beginning of the school year 1979-80 two experimental groups were formed, one in York and one in Leeds. Each experimental group had a matching control group, in which the pupils were aged 13+ and would have to choose during the year whether or not to continue with French in the next year. All the pupils completed an attitude questionnaire during the first two weeks of the first term. After this, the experimental pupils were given a document describing the graded objectives and tests, and the certificates which they could win. The parents and teachers of the two groups also completed attitude questionnaires, at the same time. Later in the year, both groups of pupils took a test of aptitude for foreign language learning. Space does not permit a detailed account of the responses to the pupils' attitude tests: these are given in the report and make instructive and fascinating reading for language teachers. What is extremely significant is that, in both of the tests they took, the two groups were very closely matched: at the beginning of the year the attitudes and aptitudes of experimental and control pupils were almost identical. The aptitude test results also showed that the language learning ability of the experimental pupils reflected extremely well the full range of ability in comprehensive schools.

The parents' questionnaires also present a lot of invaluable information for language teachers, but from the purely experimental point of view what matters is that the two groups of pupils started the year with parental support at the same level. A similar picture emerges from the analysis of the teachers' questionnaires. All the evidence indicated that the experimental and control groups were very closely matched at the start of the school year.

FL013017

At the end of the school year, pupils, parents and teachers completed the attitude questionnaires again. This was done after the experimental pupils had taken a graded test, had learned the results and, where appropriate, been awarded a certificate. The results of this second administration of the questionnaires were most striking and completely substantiate the claims made for graded objectives and tests by their supporters. Very briefly, the results of the pupils' questionnaires can be summarised as follows:-

1 At the end of the year, the experimental pupils showed very significantly more positive attitudes to learning French than the control pupils.

2 The results in Leeds and York were very similar, but the gains were more marked in the York schools. This was the first year graded tests had been used in Leeds and the third year they had been used in York. This clearly suggests that the gains from this approach come quickly and that they persist, and even improve with time - perhaps as confidence and experience grow.

3 The more positive attitudes of the experimental pupils were shown equally at all points of the aptitude scale.

4 Experimental pupils who failed a graded test proved to have attitudes as positive as those who had passed.

5 Significantly more experimental pupils expressed an intention to continue with French the following year and reports from the schools indicate that they did in fact continue.

There is a detailed examination of the results in the report and a discussion of some of the possible implications for the classroom. In a country where surveys regularly show boys as having attitudes to foreign language learning which are far less positive than those of girls, all foreign language teachers will surely be encouraged by the great improvement in the attitudes of the experimental boys as compared with those of the control group boys. It appears that the attitudes of the boys who took a graded test were as good as those normally held by girls at the same stage. Teachers who have not yet introduced a system of graded tests can also be encouraged by the knowledge that the evidence indicates that they could adopt the graded syllabuses and tests produced in Leeds and York and reasonably expect to achieve gains similar to those noted in the experimental group. As far as the pupils are concerned, the report concludes that "the introduction of graded syllabuses and tests seems to bring considerable gains and no losses".

The analysis of the parents' replies shows a trend similar to that of their children: while the control group parents' responses remained unchanged over the year or became less positive, the experimental group parents' responses all tended to become more positive. There is a strong challenge in the report to foreign language teachers who blame their pupils' negative attitudes on parental apathy or hostility. At the beginning of the year, some 80% of the parents indicated that they wanted their children to continue with French and a similar number said that they were glad that their children were learning French. The report makes the suggestion that the view generally held until now, that pupils' attitudes are largely determined by their parents, may be over-simplistic. It suggests that the attitudes of pupils and parents influence each other and that the views held by parents may well be influenced by their children. This would imply that teachers are very instrumental in forming the attitudes of their pupils towards foreign language learning and, through the pupils, their parents. Teachers will no doubt discuss the implications of this power and responsibility, and they should be helped in this by the suggestions in the report.

Compared with the numbers of pupils and parents involved, the number of teachers who took part in the evaluation was, inevitably, relatively small. Any interpretation of their replies must, therefore, be tentative. Both groups of teachers seemed generally well disposed towards graded syllabuses and tests, and their views remained stable over the year. They believed that graded tests helped lower ability pupils to achieve higher standards and that they also benefited higher ability pupils. They also reacted in a positive way to the effects of the approach on teaching courses and methods. There were no strong objections on educational grounds.

The teachers' replies do, however, provoke a number of questions, some of which suggest problems which may arise in the future from the very success of graded objectives and tests. For example, a minority of teachers supported the view that it is better to concentrate on teaching French to higher ability pupils. As graded tests appear to encourage a wider range of pupils to continue longer with French (some 62% of the experimental pupils chose to continue after options as compared with the national average of 33%-35%) this could present those teachers with a painful dilemma: their success with a wide range of pupils before options throws up a challenge after options. Judging by the discussions provoked amongst teachers in Leeds and York by this part of the report, teachers elsewhere will be helped by it to understand their own attitudes more clearly and to draw up sounder policies for future developments in their own schools.

The last chapter of the report was written by the teachers involved in the experiment. Before writing it, they read the rest of the report and met in two groups, in Leeds and York, to discuss their reactions to it. In this chapter, the teachers expand on the statistical findings, offer explanations for the results based on their own professional experience and respond to some of the challenges referred to above. They also raise a number of interesting questions and challenges of their own! This is an invaluable chapter and it evokes in graphic ways the transformations produced in classrooms by the introduction of graded objectives and tests, and the inevitable changes in methodology. One teacher tells how one recalcitrant pupil framed her certificate and secretly displayed it in her bedroom alongside her Sid Vicious poster! What comes over very strongly in the teachers' comments is a feeling of great confidence in themselves, their pupils and the future success of foreign language learning in their schools. It is clear that this confidence has not always been there and that their scheme of graded tests are very largely responsible for it. Their response to the suggestion that the success of graded tests may present some teachers with the "painful dilemma" referred to earlier is a good example of their impressive and thoughtful confidence: "The feeling was that, at present, we can and should cope with a substantial majority of children studying a foreign language to the age of 16, and hope that, as teachers' knowledge, expertise, confidence and experience grow, this will gradually increase, possibly eventually to cover all children. Clearly, teachers foresaw problems in the case of pupils with 'remedial' problems and of very slow learners. But we felt it pays to remember a parallel change of attitude and expectations which has already come about over the last decade, concerning the earlier years of secondary education (11-14)".

The developments in graded objectives and tests are starting to have a marked influence on public examinations at 16+. Several groups of teachers in different parts of the country have submitted their Level 4 syllabuses to their local Certificate of Secondary Education (CSE) Boards, and these have been accepted. The expected clash between a norm-referenced approach can, it is hoped, be avoided. A working party of the Joint Matriculation Board is currently working on an approach which could replace the existing CSE, and GCE examinations in French, and which could offer schools the flexibility they need to meet the needs of the wide range of pupils who are now starting to continue to study French until the age of 16+. In this model, there would be clearly defined objectives for listening, speaking, reading and writing at two levels, provisionally called Basic and Extended. The compulsory "core" for all candidates would be basic listening, reading and speaking (it being understood that these labels do not imply a restriction to single skill tests and that the speaking test, for example, will include some listening in a test of communication where the emphasis will be on speaking). Success in these three basic tests would earn a Grade 6 (on a scale 1-7) and success in two of them would attract a grade 7. "Success" would be defined as a high level of performance, and a minimum pass mark in each test would probably be 60%. Candidates would take whatever additional tests from this range that their teachers considered appropriate; for each one successfully completed, the candidate would move up one grade. So, a candidate who reached the required standard in six of the eight tests would receive a Grade 3. In addition, his certificate would show exactly how he earned this grade by putting asterisks in a grid, e.g.

	Listening	Speaking	Reading	Writing
Extended	*	*		
Basic	*	*	*	*

This, combined with a clear description of what was meant by listening, speaking, reading and writing at each level, would make clear to all "consumers" (pupils, parents, employers, universities, etc.) what each candidate could do successfully. It would also mean that each grade was a positive assessment, rather than the negative thing anything other than a Grade A is in the present system: even a Grade 7 would only be gained if the candidate really could do what was required to get asterisks in two boxes. In the present system, all grades below Grade A are negative grades: they only show that the candidate was not as good at something(s) as candidates who obtained a Grade A.

This approach has also been considered by the working party of the GCE and CSE Boards' Joint Council for 16+ National Criteria and has attracted some support. This working party was persuaded that high standards of performance are much more likely if the syllabus defines clearly and precisely what will be expected of candidates in all skill areas and at all levels. For each skill area and at each differentiated level it was agreed that a syllabus should define the tasks to be performed, topic areas to be covered, vocabulary, structures and functions to be used either productively or receptively and the methods of assessment to be adopted.

When discussing suitable techniques of assessment, it was agreed as a basic principle that the tasks to be set within an exam should be authentic and valuable outside the classroom. It was also agreed that, in principle, material presented to candidates should be authentic French, whether spoken, written or printed. These principles at once raised doubts about the future of such traditional examining techniques as multiple-choice questions, dictation, prose translation and reading aloud.(1)

For many years, foreign language teachers in the U.K. have blamed their problems on public examinations at 16+, often with good cause. Such examinations clearly do have a backwash effect, as was recently confirmed by the Secretary of State for Education in "The School Curriculum" (DES, 1981). It has to be accepted that the effects of testing on teaching have often been negative and that objectives and testing techniques have too often been determined because they suited testers, rather than because they encouraged good teaching and learning. This has, not unnaturally, bred much distrust amongst many teachers towards tests and testers, and testers have not always helped to prevent this: certainly, the tendency of some testers to respond to such concerns by setting up a smoke screen of jargon and statistics (lies, damned lies and ...) or by condescending paternalism, ("Well, of course, you teachers can't expect to understand the scientific intricacies of testing, so just leave it all to us ...") has strengthened this scepticism amongst teachers. Given effective co-operation between testers, teachers and learners, the development of objectives and tests can have an enormously positive effect on learning. This can only come about, however, if the testers realise that they have as much to learn from the teachers and the learners as they have to teach. The evidence certainly supports the view that co-operative work along these lines in the elaboration of graded objectives and tests in the U.K. has given new life, new insights and new enthusiasm to learners, teachers and testers.

Note

1 A French exam, based on these principles and containing only authentic French, has been produced by a working party of teachers in York, with the help of the Institut Français in London. It is hoped to try this out in the near future.

ED217728

FL013018

THE SPANISH ATTRIBUTIVE CONSTRUCTION SER/ESTAR + ADJECTIVE : REAL USE PHENOMENA SUBMITTED TO ADULT LEARNERS.

N. Delbecque and J. De Kock (Katholieke Universiteit Leuven)

1. Spanish is said to have two main attributive verbs, ser and estar, whose respective use causes some difficulties for the foreign learner. The criteria employed in most manuals to differentiate ser and estar followed by an adjective, are mainly semantic. The comments devoted to the problem in the handbooks intended for foreign learners receive a lot of space, out of proportion with respect to the remaining (true!) grammatical problems.

Most exercises offered to foreigners contain a relatively too high proportion of estar-constructions in comparison with real usage. A quick check of seven manuals (not chosen specifically for this purpose) shows that the exercises focusing on the choice between ser and estar before an adjective present from 32.6% up to 58.4% cases of estar, i.e. an average of 44.9% (1).

Let us turn now to real use data. The test corpus consists of the ser/estar + adjective constructions appearing in sixteen modern prose texts : essays, articles, discourses and conferences of modern and contemporary authors, collected in De Kock et al. (1979) (81,158 occurrences, 8,504 verbal nuclei). We count 403 verbal nuclei with ser or estar (with a total of 426 adjectives - some juxtaposed or coordinated) (2). For practical reasons we omitted the three last occurrences, considering thus 400 cases; 376 of them have ser, only 24 - i.e. 6% - have estar (3). These figures and a careful examination of the examples allow us to conclude that there is a disproportion between the attention generally paid to the opposition between ser and estar with qualificative adjective in the teaching of Spanish as a foreign language and the manifestations in our corpus.

The presence of a whole bunch of exercises with examples of estar - an exaggerated proportion - leads to a false idea in the learner's mind about the probability of encountering estar plus adjective, and explains his overgeneralization of the use of estar.

The following experiment verifies to what degree foreign students deviate from the native usage. At the same time it represents an attempt to replace the usual exercises by others which are much nearer to reality as far as the relative frequency of the alternatives is concerned.

2. The test is made up of the 400 cases already mentioned, reproduced in the arbitrary order of their appearance in the texts - following the alphabetical order of the names of the authors; the test-items are accompanied by the surrounding context in order to facilitate understanding and a blank space signals the place of the verb (1). The student is asked to complete with ser or estar (without conjugating), to put only one verb even if both seem possible and to fill in all the blanks.

The test-items were submitted to five groups of Flemish students from 19 to 23 years old, at university level : two groups are students of the first graduate level of Romance Philology, after two years of Spanish classes at a rate of two hours a week (I and II). Two other groups belong to the third year of an institute for translators and interpreters, having had Spanish classes for two years some eight hours a week (III and IV). The last group is made up of students of the second year of Romance Philology, after one year of Spanish classes of only two hours a week (V) (4). The experiment consists in verifying to what degree the students fill in the original verb, in which cases they correspond, in which they do not.

Table 1 : Relative frequency in percentages of the successful answers according to the group and the series of examples.

	I	II	III	IV	V
1-100	(5)	79.7%	88.8%	84%	73.9%
101-200	77%	78.8%	84%	90%	68.7%
201-300	87%	78.3%	88.5%	91%	66.7%
301-400	91.6%	75.7%	84%	9[illegible]%	66.5%
global percentage	85.2%	88%	86%	89%	69%

The figures show that the students of the second year of learning (V) obtain, whatever their background, a percentage of good guesses inferior to that of the more advanced groups. The results of the students with classes two hours a week (I and II) are inferior to those of the students with eight hours in the same year (III and IV); with the same grammar teacher the results vary according to the number of class hours (II and III). In the groups which have solved the totality of the test-items the variation in the results of the four series is less than 5%.

This means that the variations observed among the series and the groups are mainly due to reasons exterior to the problem itself; they are characteristic of the testees.

Considering the absolute frequency of the unsuccessful answers and their distribution (6), we see that the majority, an average of 87.5% (from 78.3% for the series 301-400 up to 98.2% for the series 101-200) are due to overgeneralized use of estar.

If we examine the relative proportion of unsuccessful answers, depending on the expectation of ser or estar, it appears that due to the use of ser being 9 times more frequent than that of estar the rare cases of estar are the ones which receive relatively more unsuccessful answers. The unsuccessful answers in cases where ser was expected appear regularly throughout the corpus, whereas ser inaccurately substitutes estar in a few very marked cases.

Even without applying the X^2 test, it can be ascertained that the number of successful and unsuccessful answers varies significantly according to the verb expected (ser or estar), and according to the series and the group. This means that the unsuccessful answers are not distributed arbitrarily in the whole corpus, and that it is necessary to look for the explanation in the accompanying adjectives and the surrounding context.

3. From the outset, the use of ser or estar could be related to three formal factors : the grammatical category of the subject, the presence of complements of the predicative adjective, and the form of the adjective itself.

When the subject is an infinitive the verb is always ser (49 cases) (1). Although we have a lot of good answers, it cannot be said that the solidarity between ser and infinitive-subject is really assimilated by the testees. It is a grammatical feature which needs more emphasis in teaching.

When the subject is a clause, introduced by que (45 cases) or not (8 cases) (1), the verb is also ser (except with claro). Even without considering the exception, we observe a number - though small - of incorrect answers. The fact that the students do not recognize this type of grammatical link in one hundred percent of the cases reveals that here they disregard the form.

The presence of prepositional phrases depending on the predicative adjective does not have any privileged relationship, neither with ser nor with estar (7).

In the examination of the items we can, thus, disregard the formal features just mentioned and focus our analysis on other contextual elements and on the adjectives themselves.

4. If we consider only the number of unsuccessful matches per sample, there is no way of discriminating between unsuccessful answers proper to the group in its entirety and those particular to part of the group, nor can we see the impact of the great number of cases with a moderate number of unsuccessful answers.

In order to discriminate the results according to the testees and the success, it is convenient to take into account the distribution of the correct and the incorrect matches in the groups. We evaluate the importance of the number of successful or unsuccessful answers for each item in particular by taking into account to what degree the answers come from testees who succeed often (scoring above the average) or rarely (scoring below the average). The fact of making a good or a bad match for an item is the less significant the more it is distributed equally among all the testees. Therefor we made use of the "exact Fisher-test", which allows us to analyze discrete, dichotomous (and not too numerous) data as the ones we have (8).

The test determines whether the two groups - respectively the testees scoring below or above the average - can be differentiated by the proportion with which they fall in both categories - good or bad matches. This way automatic computation (9) attributes to each item a probability value between zero and one: one means that the observed distribution occurs with a very high probability. This indicates either that nobody or everybody succeeds, or that successful and unsuccessful answers are distributed equally among all testees. Zero means that the observed distribution has a very low probability of appearance. This indicates that the unsuccessful answers come unilaterally from one of the two groups, generally from the participants scoring below the average.

5. Figure 2 presents the 29 items with the highest Fisher-value (one). The high frequency of successful answers shows that here

well assimilated uses are at issue, with current and little diversified adjectives and subject in verbal form (10). Seven adjectives however are striking, because they received a lot of unsuccessful answers from all testees : *loco*, *sincero*, *vivo* and *consciente* - which can be related to different possible interpretations (11) - , the particular case of *claro*, *clarisimo*, and *"vien"* (in a quotation).

The items with a low Fisher-value are the ones which interest us most : they need particular stress in teaching - a new approach maybe. The examples which are appropriate to discriminate between subjects comparable to the testees are extracted by their high number of unsucessful answers, distributed unequally throughout the sample. (Those are supposedly less understood).

2. Fisher-value : 1 : (1) references, (2) verb, (3) percentage of successful answers, (4) adjective, (5) category of the subject (*P* = clause, *que-P* = clause introduced by *que*, inf = infinitive; *todo* and *esto*), (6) contextual elements *cf. infra*.

	(1)	(2)	(3)	(4)	(5)	(6)
4	AYA081	estar	87	claro	P	
16	AYA371	ser	76,5	claro	queP	
30	BEN145	ser	97,6	util	inf	
34	BAN195	ser	98,8	latinoamericano		
53	CAR227	ser	88,2	oblomovista y apolítico		si...
77	CAS315	estar	76,5	"vien"	inf	allà
89	CAS429	ser	100	necesario	inf	
97	FER084	estar	52,9	loco		
111	GAL169	ser	100	necesario	queP	
117	GAL473	ser	98,5	necesario	queP	
119	GAL485	ser	97	necesario	inf	
122	GAL496	ser	71,2	sincero	inf	
151	GA0295	ser	98,5	probable	queP	
152	GA0308	ser	91	improbable	queP	
196	JIM176	ser	100	lógico	P	
219	JIM611	estar	57,5	vivo		siempre
242	MAR401	ser	96,5	cierto	queP	
265	ORT314	ser	96,5	evidente	queP	para la conciencia contemporánea
269	ORT351	ser	79,3	bueno	queP	
280	PER009	ser	51,7	clarísimo	esto	a poco que...
291	PER213	ser	95,4	mejor	inf	
294	PER283	ser	97,7	cierto	queP	no
296	PER335	ser	100	posible	inf	cuando... ya...no
298	PER339	ser	95,4	uno y lo mismo	todo	
299	PER340	ser	100	posible	inf	no
336	VAR023	ser	86,8	impreciso		
350	VAR170	ser	100	evidente	queP	para mi
393	VAR621	ser	100	latinoamericano		sin dejar de...
399	VAR687	estar	64,5	consciente		también

For 185 items - nearly half of the 400 cases examined - the value obtained is higher than 0.4; in 100 other cases the Fisher-value falls between 0.1 and 0.4; 138 other items fall between 0.01 and 0.1; and the 77 remaining cases receive a Fisher-value below 0.01.

In all of these 77 cases with Fisher-value below 0.01, as can be seen on figure 3, the testees did not succeed because they chose estar instead of ser. This confirms that there exists a clear tendency to overgeneralize the use of estar plus qualificative adjective.

3. The 77 items with Fisher-value below 0.01 in rising order (1), their reference (2), the verb (3), the adjective (4), the subject (5), other restrictive contextual elements (6).

(1)	(2)	(3)	(4)	(5)	(6)
0.000057	303	SER	endémico,permanente		
0.000072	340	SER	idealista		si ...
0.000087	283	SER	razonable	esta manera	que es como si no existiera
0.000094	284	SER	facultativo		si...
0.000095	38	SER	lamentable		se me ocurre que
0.000111	11	SER	sencillo		si no lo fuera
0.000151	64	SER	atribuible	la falta de reacción	no
0.000173	49	SER	ilustre		
0.000242	264	SER	individualista,sino personalista		no
0.000263	201	SER	externo,superficial, disperso		
0.000326	12	SER	enrevesado, burlesco		
0.000349	352	SER	separable		no
0.000495	307	SER	prematuro		(después de una interrogación)
0.000627	233	SER	unánime		
0.000819	398	SER	parejo	la incidencia	
0.000871	121	SER	vasto,complejo		
0.000917	362	SER	decisivo	lo irracional	
0.000929	185	SER	inocente		no ... jamás
0.000938	29	SER	despreciable	una conducta	
0.000938	1	SER	disparatado, sino lógico		en tal coyuntura no
0.00126	95	SER	inocente		según se nos describe
0.00139	182	SER	completo	la descripción	no... nunca... si...
0.00139	184	SER	completo	la descripción	nunca... si...
0.00169	42	SER	bueno		... aunque
0.00185	377	SER	admirativo		... cada vez menos... más
0.00193	387	SER	manifiesto		a condición de...
0.00200	147	SER	sumo	frecuencia	... en casos...
0.00203	310	SER	inviolable, indiscutible		
0.00231	232	SER	desagradable o poco favorable		
0.00231	259	SER	impropio		
0.00231	249	SER	inepto, y a menudo grotesco		hoy
0.00263	50	SER	lamentable	la situación	pese a que
0.00271	331	SER	cerebral		cuando leo
0.00271	325	SER	consabido		aunque
0.00272	158	SER	imprevisible	los efectos	actualmente
0.00296	236	SER	laudable		desde el punto de vista
0.00301	114	SER	adverso		
0.00304	66	SER	otro	la situación	en el siglo XVII
0.00307	359	SER	espontáneo		
0.00334	235	SER	increíble		desde otro punto de vista
0.00384	31	SER	grave		acaso (interrogación)
0.00387	195	SER	aristocrático, o popular		
0.00387	103	SER	capaz		todavía
0.00400	290	SER	imprescindible		tal es el caso / por... que
0.00400	281	SER	lícito,probable		supuesto que... (interrogación)
0.00439	110	SER	cordial		
0.00439	332	SER	eminente		ni
0.00472	79	SER	nuevo		mirado al día de hoy, no
0.00495	180	SER	coherente		a condición de
0.00500	203	SER	distinto		aunque... según...
0.00527	22	SER	solidario		
0.00592	10	SER	cierto		si
0.00616	293	SER	débil		luego, a medida que

(1)	(2)	(3)	(4)	(6)
0.00636	250	SER	forzoso	
0.00640	129	SER	exigible	en esta hora
0.00684	115	SER	adverso	
0.00684	116	SER	consecuente	solo así
0.00719	240	SER	flojo	acaso no / en nuestras letras
0.00726	155	SER	discutible	se trataría
0.00726	183	SER	intenso	siempre más
0.00726	162	SER	peor	
0.00726	131	SER	propicio	puedan (subj.)
0.00742	285	SER	estéril, inepto	si...
0.00742	329	SER	peor	debería (cond.)
0.00742	284	SER	razonable	no
0.00754	309	SER	hacedero	por ahora no
0.00756	86	SER	indisoluble	
0.00842	160	SER	exigente, recompensado	(interrogación)
0.00846	73	SER	inseparable	
0.00850	357	SER	arbitrario	
0.00850	302	SER	escasísimo y secundario	como no sea
0.00887	98	SER	otro	
0.00895	206	SER	nuevo	por... que sea. Sólo con volver a.
0.00908	153	SER	difícil, fácil	si
0.00955	158	SER	superior y más importante que	están convencidos
0.00967	378	SER	incompleto	a veces
0.00972	373	SER	absoluto	ya no tan

The question arises as to the possible reasons which induced the testees to choose estar instead of ser and to which degree those motivations originate from their former grammatical training.

6. The rule here is usually reduced to an opposition between "transitorio", "desinente" or "accidental" for estar, and "permanente" or "inherente" for ser. The quoted terms are the ones of the Real Academia Española (1976:3.3.3.b). They are similar to the ones found in other grammars, manuals or comments, where they are generally not accompanied by the reserves which the Spanish Academy itself puts forward (12). Some contextual indications possibly provoke the overextended use of estar with respect to ser, when they evoke circumstances or conditions which are transitory, incidental of subjective in nature. They are likely to prevent the quality expressed by the qualificative adjective from being considered permanent or inherent.

In the light of this experiment it appears that such circumstantial elements - even when they suggest a restriction without it being effectively there - disorientate the foreign student. In most of the 77 cases with Fisher-values below 0.01 we find contextual indications which probably led to such an interpretation. They are situated not only in the adjective itself - in its morphology - but also in subjects, prepositional phrases, adverbial locutions, adverbs and subordinate clauses which accompany the verbal nucleus, in isolation or together.

Among these 77 examples with abusive choice of estar we find 18 adjectives ending in -a(i,u)ble, 7 in -ente, 3 in -at(is)ivo, 3 in -a(i)do, 2 in -ista and 2 in -esco. Adjectives occurring more than once are : (in)completo (3), adverso (2), nuevo (2), otro (2) and peor (2).

The 18 adjectives in -a(i,u)ble with Fisher-value below 0.01 are : razonable (2), lamentable (2), atribuible, separable, despreciable, inviolable (indiscutible : with the same verb), desagradable (o poco favorable : with the same verb), imprevisible, laudable, increible, imprescindible, probable, exigible, discutible, indisoluble, inseparable.

In the corpus these and the other adjectives in -a(i,u)ble - occurring in 67 examples - never appear with estar The testees often fill in estar, probably being misled by the suffix. It expresses a "passive potential aspect" (Pottier 1968:124) which they relate to the transitory, teminative or accidental meaning of the rule they learned.

In order to know to which degree this property can be considered responsible for the result, it is necessary to know how many times this type of adjective occurs without being associated with a low Fisher-value. Less than the fifth part (13 cases) lead to a value higher than 0.4 - they are generally very current adjectives such as (im)posible, (im)probable, responsable. A value between 0.1 and 0.4 corresponds to 12 cases, a value between 0.01 and 0.1 to 24 cases, and a value below 0.01 to the 18 cases just mentioned.

Next to this series, the 77 propositions with Fisher-value below 0.01 present 7 cases with adjectives in -ente : coherente, consecuente, eminente, exigente, inocente (2), permanente (13). On the other hand, the corpus contains two adjectives in -ente in a construction with estar : consciente and presente. In the great majority of the remaining cases the Fisher-value does not exceed 0.4. Here again, it can be supposed that the ending -ente, which expresses "virtuality", suggests to the testees an interpretation of "non-permanent", "not-inherent", and, consequently, causes them to select estar.

The fact that the adjectives in -ante (apremiante, estimulante, interesante, flagrante, motivante, brillante, importante) - with verbal stem - all appear in constructions with ser and equally receive a relatively low Fisher-value - not one above 0.25 - points in the same direction.

Analogous explanations and interpretations can probably be adduced for the selection of estar with adjectives in -ivo (admirativo, decisivo, facultativo) and in -ista (idealista, individualista), and the same holds true for stray adjectives as peor, hacedero, prematuro, etc., whose meaning can only difficultly be made to coincide with "permanent" or "inherent".

7. Going back to the mere frequencies, let us consider now the 80 items which receive a successful guess in less than 75% of the cases. This series - represented in table 4 - corresponds partly to table 3 devoted to the low Fisher-test results. If an item reappears here, it means that it is characterized at the same time by a high percentage of bad guesses and an unequal distribution of good and bad matches both within the most and the least successful of the testee-groups.

Contrary to table 2, this one does contain estar-constructions. Not only a few, but the great majority : 20 of the 24 (14). In most cases they, moreover, receive a very high Fisher-value. This indicates that the misguessing is generalized, i.e. distributed equally in the whole group. The rare uses of estar do not discriminate among the students.

For the erroneous selection of estar instead of ser, on the other hand, there is, in most cases, some restrictive indication which can be held responsible for it (15).

4. 80 items with less than 75% successful answers

less than 75%		adjective	contextual elements	fisher-test
1		disparatado, lógico	en tal coyuntura no	0.0009
10		cierto	si	0.5059
16	ESTAR	claro		1.0000
22		solidario		0.0052
24		injusto		0.0188
26	ESTAR	libre	antes	0.6602
27	ESTAR	preso	ahora	0.6277
29		despreciable	(una conducta)	0.0009
37		capaz	estoy dispuesto	0.5750
38		lamentable	se me ocurre que	0.0000
44		lamentable		0.0337
50		lamentable	pese a que (la situación)	0.0026
56		válido	por el momento	0.2306
64		atribuible	no(la falta de reacción)	0.0001
65		visible	aún	0.0459
66		otro	en el siglo XVII (la situación)	0.0030
68		sucio		0.0644
70		seguro		0.0682
71	ESTAR	claro		0.8269
75	ESTAR	'quedo'		0.3180
76	ESTAR	seguro		0.3703
77	ESTAR	'bien'		1.0000
78		claro, limpio		0.4810
79		nuevo	mirado al día de hoy, no	0.0047
92		análogo		0.6181
97	ESTAR	loco		1.0000
99		objetivo, comedido		0.3750
110		limpio, puro		0.0888
114		adverso		0.0030
115		adverso		0.0068
116		consecuente	sólo así	0.0068
126		superable		0.5937
129		exigible	en esta hora	0.0064
133		excesivo	en momentos	0.4647
139	ESTAR	presente	en su jornada	0.2071
141		enfermo		0.8062
143		insólito		0.0312
154		superior		0.0508
158		imprevisible	actualmente (los efectos)	0.0027
160		exigente, recompensado	(interrogacion)	0.0084
167		infiel	en ello	0.1415
170		absurdo	dicha situación	0.0165
177		doloroso	no...siempre	0.0117
182		completo	no...nunca...si (la descripción)	0.0013
184		completo	nunca...si (la descripcion)	0.0013
190		subjetivo, irracional, arbitrario	a un tiempo	0.4672
206		nuevo	con volver a...	0.0089
207		claro	con volver a...	0.5462
214		claro		0.6471
219	ESTAR	vivo	siempre	1.0000
230		lesivo, ponzoñoso	que suele ser benévolo	0.0181
234		canjeable		0.3292
238	ESTAR	interdicto	por el gobierno	0.3878
240		flojo	acaso/en nuestras letras	0.0071
251		exacto	por esta vez/en parte	0.0118
266		anterior		0.0158
267		anterior		0.0158
268		bueno	sòlo definitivamente ...cuando	0.1426
275		sano	no...nunca	0.0501
278		débil		0.0287

less than 75%		adjective	contextual elements	fisher-test
280		clarísimo		1.0000
292		fuerte	a medida que	0.0124
293		débil	luego, a medida que	0.0061
309		hacedero	por ahora, no	0.0075
315	ESTAR	absorto		0.6446
319	ESTAR	bueno		0.0295
320	ESTAR	seguro	ya	0.8173
322		externo		0.0173
327	ESTAR	mejor	acaso	0.0892
328	ESTAR	claro		0.0440
[illegible]		claro		0.2049
345	ESTAR	claro		0.0215
346		inválido	desde cuando	0.1033
349		militar	ahora	0.0440
363		clarísimo		0.6365
366	ESTAR	seguro		0.4582
371	ESTAR	decrépito	pone de moda	0.0226
378		incompleto	a veces	0.0096
399	ESTAR	consciente		1.0000
400		completo	nunca	0.0750

The 10 occurrences of claro are striking in this list (32). As an interjection, "claro esta" appears to be a fixed locution (33). In the other constructions with an NP subject or a subject-clause the usage fluctuates between ser and estar, as if they were equivalent. Due to this very fluctuation it is not surprising that the presence of claro has led to a great many unsuccessful answers and, in general, to a high Fisher-value.

Besides claro, three other adjectives appear both in constructions with ser and estar : bueno, mejor, seguro (16). It is the usage of estar which figures in the list of the items with the greatest number of unsuccessful answers.

8. This confrontation of foreign students with real-use data shows that they tend to overgeneralize the use of estar with qualificative adjective. The meaning of adjectives like subjetivo, irracional, arbitrario (190) or externo (322), for instance, disorientates them. The presence of restrictive contextual elements often induces a bad guess because of the theoretical relation generally established a priori between estar and limitative referential elements.

In conclusion, it would be convenient to draw the students' attention to three main points: (a) - except with claro - Spanish never uses estar when the subject is an infinitive or a clause; (b) in spite of a 'temporal', 'accidental', 'terminative', or 'transitory' interpretation - often based on contextual indications - the use of ser is maintained in 94% of the cases; (c) adjectives ending in -a(i,u)ble, -ente (except presente and consciente), -ante, -is(at)ivo, and -ista appear in constructions with ser.

Notes

1 Detailed references are available to anyone interested.

2 Cf. De Kock (1981:149-153).

3 Exhaustive and systematic examination of other corpuses reveals analogous proportions.

4 Each student of group II and group III filled in all 400 items. The other three groups were split up in four subgroups, so that each student answered one series of 100 items (1-100, 101-200, 201-300, 301-400).

5 The cell corresponding to group I, series 1-100, is empty, because this particular test group studied large parts of the material from which the first series of examples is taken during the previous year.

6 Detailed frequency tables were omitted due to lack of space.

7 This occurs 4 times with *estar*, 57 times with *ser*, i.e. with the same proportion as the *global* use of both verbs. Prepositional phrases do not have any influence on the result : in each case there are unsuccessful matches.

8 The probability is given by the hypergeometric distribution and corresponds to the formula

$$\frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{N!\ a!\ b!\ c!\ d!}$$

where a and b represent the number of, respectively, the unsuccessful and the successful answers coming from testees scoring above the average; c and d represent the number of, respectively, the unsuccessful and successful answers coming from testees scoring below the average; and N respresents the total number of answers.
A figure may clarify the distinctions made :

	-	+	
Group I : H	a	b	a + b
Group II : L	c	d	c + d
	a + c	b + d	

Group I represents the students scoring higher than the average (H), group II those scoring lower (L). The columns identified by -/+ correspond, respectively, to the categories of unsuccessful and successful answers to each item.

9 The program was put at out disposal by F. Loosen, and applied by the computing service of the university.

10 *Necesario*, *claro*, *cierto*, *evidente*, *latinoamericano*, *posible*, *(im)probable*, with as subject 8 infinitives, 10 subject clauses introduced by *que*, and 2 without conjunction.

11 It is known that a series of adjectives is combined with *ser* or *estar* according to the point of view of the speaker, so that the unsuccessful matches - distributed equally among all - can be interpreted neither as good discriminators in the test nor as specific difficulties for foreign students because they cannot be reduced to one grammatical rule.

12 "Las reglas basadas en las oposiciones entre *permanente* (ser) - *transitorio* (estar), *permanente-desinente*, *inherente-accidental* y otras, tienen utilidad práctica indudable, pero limitada, y se prestan a interminables discusiones sobre el sentido con que deben entenderse los términos de cada pareja" (Real Academia Espanola 1976:33).

13 For the 20 remaining items with an adjective in *-ente* we obtain a Fisher-value higher than 0.4 in only 6 cases : *conveniente*, *evidente* (3), *pertinente* and *suficiente*. The others are : *corriente*, *diferente*, *eficiente*, *estupefaciente*, *evidente* (2), *exigente*, *frecuente*, *impertinente*, *improcedente*, *inconsistente*, *indiferente*, *permanente*, *suficiente*.

14 Of the 4 remaining cases with *estar*, 3 have *claro* (the numbers 4, 221 and 253, incidentally with more good guesses), the last one *lleno*.

15 In fact, most of the temporal indications (some 30) coincide with the use of *ser*, as, for instance, in :

> Cada empleado administrativo podrá continuar sustentando su propio y personal credo político, pues *en esta hora sólo es* exigible eficiencia y lealtad administrativa (129, GAL 625)

Although *estar* appears 5 times in contexts which are temporally delimitated, we also find the following paradoxical case :

> La fuente *está* siempre viva, sangre latente, aunque no se sepa (219,JIM 611).

Here *siempre* co-occurs with *estar* in a construction where the choice of the verb depends on the meaning attributed to the adjective : "ser vivo (rápido, inteligente), estar vivo (vivir, gozar de vida)" (Real Academia Espanola 1976:367).

16 If the feature (+ human) can be considered as pertinent for the subject of *estar seguro*, in the remaining cases, as with *claro*, it is a matter of marginal variation.

References

De Kock, J., R. Verdonk, R. and Gómez Molina. 1979. *Gramática española : enseñanza e investigación*, t. II : Textos. Leuven : ACCO.

De Kock, J. in collaboration with C. Gómez Molina. 1981. *Notas a la gramática española para la enseñanza en el extranjero*. Leuven : ACCO.

Pottier, B. 1968. *Lingüística moderna y filología hispánica*. Madrid : Gredos.

Pottier, B. 1971. *Gramática del español*. Madrid : Alcalá.

Real Academia Española. Comisión de gramática 1976. *Esbozo de una nueva gramática de la lengua española*. Madrid : Espasa-Calpe.

ED217729

TOWARDS A PRACTICAL MODEL OF A COMMUNICATIVE COMPETENCE TESTING PROGRAM

Robert A. Fischer (Southwest Texas State University)

References to broad principles and specific techniques of testing linguistic competence and communicative competence abound in the literature on language testing. Yet, despite the plethora of theoretical studies and experimental reports, the relationship between the two remains relatively unexplored. In fact, most of the relevant research gives the impression that linguistic competence and communicative competence are fundamentally distinct theoretical constructs with few features in common. Within the context of language testing, Davies' (1978:215) contention that linguistic competence and communicative competence represent different points along a single language learning continuum alone offered a theoretical framework in which to view the apparent relatedness between the two. Bialystok's (1979) findings that the student's use of the foreign language to participate in communicative events leads to increased achievement in the language provides empirical support for Davies' claim. This paper will examine the relationship between linguistic and communicative competence in terms of the student's acquisition of French as a foreign language and the interface of testing for each one in the context of achievement testing. In addition, this paper will also propose a model of a testing program whose features are designed to guide the student from the acquisition of linguistic competence to the ultimate goal of effective communicative competence.

Language educators universally affirm that the goal of foreign language instruction is the ability to communicate in the foreign language. It is also generally assumed that the acquisition of communicative competence presupposes the implicit or explicit acquisition of linguistic competence. Proponents of the various models of foreign language learning prevalent today do not fail to include in their models some form of language learning that can only be properly interpreted as the acquisition of linguistic competence.

For the foreign language learner, linguistic competence may be thought of as his developing knowledge of the structure and vocabulary of the language and his ability to produce and comprehend well-formed sentences in the language. The production or comprehension of well-formed sentences involves the application or the recognition of the application of usually explicitly learned rules to isolated language samples. In this restricted sense, the student's participation in initial classroom exercises may be accurately described as rule governed behavior in which his attention is focused on the application of rules to derive correct grammatical forms (1). The development of the student's linguistic competence consists of increasing the number of morphological markers,

FL013019

syntactic structures, and lexical items to which he is exposed in the instructional sequence.

Communicative competence represents for the language learner a functional re-orientation of the knowledge and skills he has already acquired and requires his adoption of a new focus for language production and comprehension. The student faces the task of using the language to send and receive messages in concrete situations and for specific purposes. Schulz and Bartz (1975), Morrow (1979:149-50), and Johnson (1979:199-201) give broad taxonomies of communicative competence from which several general characteristics can be abstracted. First, the basis for produced or received messages in communication events is primarily semantic in nature. The semantic basis of messages may be taken to pertain to what Cherry (1978:9) has called the "information content" of messages, which involves the messages' referential function of representing real world entities and events. The communicative value of a message depends on the new information it presents and the degree to which it is readily apprehended by the participant. Thus, at a minimal level of performance, the language student must attend to the comprehensible formulation of sentences which have the weight of real world references. Second, communication in a language requires the presence of a realistic context. The presence of a concrete communicative situation enables the participant to identify the relevant sociolinguistic parameters, which include the physical environment in which the communication takes place, the social status of the addressee, and the general topic of communication. For the language learner, the social status of the communication partner should trigger the appropriate use of the sociolinguistic variables of the language. The establishment of the topic of the communication creates a set of psychological expectancies about the nature of the communication, which serves to facilitate the student's comprehension or received messages and his production of new messages. Third, participants in a communication event seek to attain specific goals. The reason a person engages in the act of communication is to obtain information or to react to a request for information. In fact, the perceived value of the participant's contributions to a communication is a function of the new information he introduces in the communication which is potentially useful to his partner (see Paulston 1970:192; Cherry 1978:14; and Johnson 1979:200). Thus, the language student is called upon to provide responses pertinent to his partner's statements and inquiries. Last, the development of a series of communicative exchanges is only partially predictable. The introduction of new information or the syntactic manner in which it is expressed serves to keep the communication open-ended. Although it may be possible to anticipate a plausible range of points of discussion as determined by the context, the uncertainty of the speaker's actual utterance or the variation in his use of communicative strategies makes the course of the communication only partially predictable (Cherry 1978:15; Kleinman 1978:158-65; and Tarone 1980:429).

Since the relationship between linguistic competence and communicative competence involves, in general, a functional re-orientation of language knowledge and skills towards a pragmatic use, communicative competence may be characterized as linguistic competence mobilized for the purpose of communication (Carroll 1968: 56; 1975:111; and Candlin 1973:57). In this view, the acquisition of linguistic competence is a necessary but not sufficient condition for the attainment of articulate or literate communicative competence. That is, the student cannot communicate effectively without knowledge of the structure of the language, but knowledge alone does not assure that he will be able to interact with speakers of that language. On the other hand, spontaneous attempts at communicating with speakers of the language with no language training is excessively frustrating and leads no father than to the attainment of meager survival communication skills. Reaching an acceptable level of communicative competence requires practice and guidance in both understanding the structure of the language and using the language in realistic situations.

The student's primary task to realize his linguistic competence in terms of communicative competence is to integrate isolated areas of knowledge and skills into larger units which are of practical use in communicative situations (see Carroll 1968:56; Zelson 1978:45; Morrow 1979:148; and Palmer 1979) (2). The student has to incrementally organize new features of the language's structure with its previously learned structures in order to increase the functional capacity of his developing linguistic competence. The key to progressing from linguistic competence to communicative competence within the confines of the language classroom resides in the student's ability to integrate molecular linguistic units into molar communicative units. (The terms "molecular" and "molar" are borrowed from Bosco and DiPietro 1970:40.)

The classroom testing program, if it is to be an effective and valid instrument of evaluation, must be congruent with the goals and activities of the instructional sequence. If the goal of the class is communicative competence, the student must be tested on his ability to communicate in the language (Schulz 1977:101; Davies 1978:224). Further, if the acquisition of linguistic competence is a necessary condition for the attainment of effective communicative competence, the student's linguistic competence must also be evaluated in order to assure that he has indeed acquired the necessary knowledge and skills to attain a level of effective communicative competence. Within this perspective, direct tests of the student's ability to communicate in the language constitute a sufficient condition for reaching the goal of effective communicative competence (3). The conclusion seems inescapable that if articulate or literate communicative competence is the goal of the class, then tests of both linguistic and communicative competence have to be included in the classroom testing program (see Jorstad 1974:247; Jones 1977:239; Davies 1978:221-2; and Morrow 1979:152).

The specification of the features of tests of linguistic and communicative competence follows immediately from the characteristics of each described earlier. In general, linguistic competence tests are tests of the student's knowledge of the language. They evaluate the student's understanding of particular rules and structures and their operation in concrete language samples. Test items that require the student to convert a sentence from one form to another, such as present to past tense or singular to plural, provide immediate evidence of his ability to apply linguistic rules to derive correct language forms.

From this primary principle, two corollaries logically follow. Linguistic competence test items are basically analytic in nature and elicit reactive responses (see Carroll 1968:50-1; Clark 1972: 111). Test items like multiple choice or error identification-correction items, which require the student to analyze and to react to stimuli present in a language sample, yield a relatively direct view of his linguistic competence. In order to answer this kind of question, the student has to identify and react to the salient features that contribute to the grammaticality of the stimulus sentence. Then, he must use his knowledge of the language to judge the grammatical accuracy of his projected response.

Linguistic competence test items typically pertain uniquely to surface forms and only rarely to deeper semantic or conceptual structures. Such surface oriented test items serve little communicative purpose. For example, testing devices such as the substitution type, which focus on the morphological properties of words, refer to categories already present in semantic structure. Translation exercises may require the student to give language specific syntactic form to underlying semantic representations but still fall far short of engaging the student in an exchange of information.

Linguistic test items have traditionally been free of contextual constraints. All of the kinds of test items mentioned here so far present only isolated sentences to the student in a testing situation. More recently, however, an impressive amount of research has investigated the effects of placing test items in continuous passages by means of various cloze procedures. All the forms of cloze testing that have been proposed so far have the same essential function of requiring the student to examine the contextual features present in the passage in order to project an acceptable word to fill in the blank created by the deleted word. Cloze testing seems to measure more than knowledge of surface forms since the successful completion of the items brings into play the student's knowledge of the semantic and syntactic properties of words that may be appropriately inserted in specific frames. Yet, contextualized items do not by themselves call on the student to produce any internally generated responses which provide information of an extra-linguistic nature and, consequently, ought to be classified as linguistic competence test items.

Linguistic competence test items admit only a narrow range of responses, often a unique response. Conventional test items like fill-in-the-blank for verb conjugations or adjective agreement or combining sentences using the correct form of a relative pronoun represent the extreme form of this kind of testing. Testing devices like cloze procedures or dehydrated sentences allow only a slightly higher degree of flexibility in the student's responses.

Communicative competence tests are necessarily tests of the student's ability to use the language in personal interactions. They evaluate the student's reaction to language stimuli and his subsequent production of an appropriate response. For example, in the early phases of achievement testing in the classroom, the student might see or hear personalized questions that require him to understand the essential propositional structure of the questions, perhaps at only a literal level, and to produce meaningful, comprehensible responses.

Communicative competence test items are basically integrative in nature and elicit productive responses. In contradistinction to the test items which focus on one aspect of a sentence, communicative test items require the student to scan the entire sentence and to attend to all of its major constituents. Test items like the personalized questions mentioned in the preceding paragraph require the student to comprehend several grammatical structures and then actively use them and others to produce appropriate responses.

In addition to determing the form of specific communicative test items, the term integration has a much more general and fundamental consequence for the testing program as a whole. It is recalled that the primary task of the language learner to progress from linguistic competence to communicative competence is to integrate molecular linguistic structures into molar communicative units (4). The primary function of communicative competence testing should be to measure the degree to which the student has organized the various language structures presented in class and is able to use them in communicative situations. Further, since the kind of tests administered in the classroom implicitly informs the student which material he should learn, it might also be extended to inform him how he should learn it. The incorporation of the concept of cyclic testing in the testing program can function to encourage the student to integrate separately learned language structures into usable communicative areas. That is, the testing program can be designed such that the linguistic and communicative components apply in a cyclic fashion. In a given testing situation the student is examined on his linguistic knowledge of the content of a specific unit and, in addition, on his ability to communicate in the language using what he has learned from previous units. In this way, the content from earlier units is continuously recycled in each classroom test. The student is informed in advance that he will be expected to use the structures from the previous unit in a communicative test and that the test items will relate directly to the thematic or cultural content of that unit (5). The linguistic-communicative cycle is then repeated for the following units. The pedagogical value of cyclic testing is that it provides the student with a systematic prompt to review earlier units for the current test since his ability to use the material from those units will figure in his score. In addition to assuring distributed practice of specific language structures, the cycle guides the student in integrating individually learned language structures from each unit into usable communicative wholes and thereby increases the functional capacity of his developing linguistic competence.

Communicative competence testing must involve a meaningful exchange of information between the student and his communication partner. As such, communicative test items should address areas of concern of which it may be reasonably supposed that the student has some knowledge or may readily obtain that knowledge in the course of the communicative situation. Testing devices, such as a questionnaire which solicits information from the student about his own interests and activities, require him to provide specific information about his own physical and psychological universe.

Communicative test items have to take into account the presence of contextual constraints perhaps introduced by prefatory statements and then elaborated within the item itself. For example, the instructor can present to the student in a testing situation a letter from an imaginary person in the foreign country, such as a letter from a pen-pal or a business inquiry, and explain the general lines of the communicative context. The letter itself will set the tone for the student's response.

Communicative competence test items admit a wide latitude of responses, if not an open-ended range of responses (see Valette 1977:66; Brière 1980:91). The number of correct responses to linguistic test items is severely limited, but the number of possible answers to the questions of the questionnaire above, for example, is virtually limitless.

A summary of the specifications of linguistic and communicative competence testing is provided in table 1.

Table 1

Specifications of Language Testing Techniques

LC	CC
1. evaluate knowledge	1. evaluate actual use
2. analytic-reactive	2. integrative-productive
3. nonpurposive	3. purposive
4. may be context independent	4. must be context dependent
5. permit narrow range of responses	5. permit wide range of responses

The distinction between linguistic and communicative testing methods is not categorical but instead resides in the degree to which a test item fulfills the specifications listed in Table 1. Linguistic test items take on more of a communicative function in proportion to the number of communicative specifications they meet. Thus, for example, a test item may be nonpurposive but still subject to contextual constraints, as in cloze testing. This kind of item represents a transistion from linguistic to communicative testing. A test item which satisfies all the communicative specifications will possess a relatively high degree of construct validity and, consequently, will be more highly valued as a communicative test item. Further, the specifications in Table 1 really amount to polar descriptive features since the pedagogical function of any language test can be described in terms of these features. As such, they may constitute evidence for the claim that linguistic competence and communicative competence are two points along a single language learning continuum.

A pilot experiment designed to evaluate the effectiveness of this kind of testing program in a classroom setting was undertaken in first year university level French classes. The experiment involved two sections of second semester French language classes and sought to isolate the method of testing as the single variable. Both the control and experimental sections were taught by the same instructor, who used the same conventional textual materials and the same teaching strategies. What may be viewed as a cognitive model of language instruction was followed for both sections, characterized by deductive presentations of language rules and structures, exercises sequenced to progress from mechanical drills to communicative interactions, and the use of English in the classroom for explanations and as translation devices. Each group was given three major classroom tests, one hour each, and a common two hour final examination. In addition, the grammatical and lexical content of the language tests given in both sections was virtually identical. The Modern Language Association Cooperative Foreign Language Tests, Forms LA and LB, were administered as pre- and post-tests. Neither group was informed of the real purpose of the pre- and post-tests, nor were the members of either group aware that they were the subjects of an experiment. The control group received traditional discrete-point tests which measured the degree to which the members of that group had mastered various rules and structures of specific lessons. The testing program of the experimental group consisted of the cyclic administration of linguistic and communicative competence tests as has been described here. The test items of each component included elements of the specifications presented in Table 1.

Comparison of the scores of the two groups on linguistic competence test items reveals that the experimental group achieved a higher degree of linguistic competence, thus corroborating Bialystok's (1979) findings that practice in using the language for communicative purposes leads to increased language proficiency. The scores from two linguistic competence testing devices of the semester final examination are summarized in Table 2. Since neither the number of subjects involved in the experiment was sufficiently large to support the assumption of normal distribution of the data nor the test items subjected to item discrimination analysis to determine the weight of individual questions, the median scores for each group are listed.

Table 2

Median Scores from LC Test Items on Semester Final Examination

Item type	Control group N = 17	Experimental group N = 17
Sentence rewriting (pronominalization exercise)	14.0	15.0
fill-in-the-blank (passé composé exercise)	19.0	23.0

The Mann-Whitney U test for sums of ranks indicated that the scores of the experimental group were significantly higher than those of the control group, $p < .05$.

The median scores from the Modern Language Association Cooperative Foreign Language Tests pre- and post-tests are given in Table 3.

Table 3

Median Scores from MLA Coop Tests

Group	Pre-test (Form LA)	Post-test (Form LB)
Control group		
N = 13		
Reading	146.5	155.0
Writing	151.0	152.0
Experimental group		
N = 11		
Reading	153.5	164.0
Writing	158.5	164.0

Again, the Mann-Whitney U test for sums of ranks indicated that the scores of the experimental group were significantly higher than those of the control group, $p < .05$.

In addition to the linguistic competence testing component, the semester final examination also included a communicative competence testing component. One of the testing devices involved a paragraph length response, of a specified minimum number of sentences, to a question relating to American university student dress standards. Two scoring procedures were used to grade the students' responses, both of which sought to evaluate literate communicative competence. The first procedure consisted of assigning a subjective global score of communicative effectiveness, which represented the appropriateness and the general quality of the information content of the student's response. The global score, ranging from 1 to 6, was multiplied by a factor of 10, and that composite figure was divided by the number of errors in the paragraph. The second procedure consisted of counting the number of intelligible clauses in the student's response, multiplied by 10, and divided by the number of errors (6). The knowledge of and the ability to use gross propositional structure and major morpho-syntactic processes were taken as an absolute minimum of performance. Since the established goal was literate communicative competence, an error was defined as a violation or misuse of the grammatical structures or lexical items that had been presented and practiced in class. No attempt was made to take into account the degree to which an error interfered with communication. The median scores for both groups are listed in Table 4.

Table 4

Median Scores from CC Test Item on Semester Final Examination

Method of grading	Control group N = 17	Experimental group N = 17
$\frac{\text{global score (1-6) X 10}}{\text{no. of errors}}$	3.75	4.84
$\frac{\text{no. of intell. clauses X 10}}{\text{no. of errors}}$	10.0	11.16

The global score method of grading did not yield significant results. However, the Wilcoxen matched-pairs signed-ranks test indicated that the median score of the experimental group obtained by the number of intelligible clauses method was significantly higher than that of the control group, $p < .05$.

Although the sample sizes in the experiment were small and the results may support only tentative conclusions, the preliminary findings suggest the general direction of testing for effective communicative competence. It would be expected that practice in dealing with communicative test items in the tests given to the experimental group during the semester would lead to an increase in communicative performance and production on the final examination. Of greater interest is that communicative competence, integrated with linguistic competence tests, led to substantial improvement in literate communicative competence, functioning as a variable dependent upon linguistic competence.

Since today's language teachers have at their disposal a variety of testing techniques and diverse kinds of test items to measure linguistic and communicative competence in the classroom, they would be well advised to critically examine the effect of potential test items in light of the characteristics of the student's developing linguistic and communicative competence and select those items that are appropriate to the intentions of the particular language test. When these pertinent factors are taken into account and the actual testing devices are integrated according to an organizing principle such as the cyclic administration of linguistic and communicative testing components, the classroom program of achievement testing can have an appreciable impact on learning and guide the student to attain the goal of effective communicative competence.

Notes

1 The application of rules to concrete language samples forms the cornerstone of cognitive methodology. Chastain (1969:105; 1976:148) claims that the conscious acquisition and application of rules is one of the first objectives of the language learning process. Rule application is the activity that bridges the two stages of Rivers' (1975:4) "Skill-Getting" tasks.

2 Palmer (1979:170) has provided perhaps the clearest statement of this essential relationship. He has referred to what he terms "compartmentalized control" as mastery of a set of discrete teaching points and "integrated control" as the use of those discrete points in communication. Further, he has empirically demonstrated that "integrated control" is the most important ability that underlies communicative competence.

3 Jones (1976:244) has presented the same argument in a slightly different form. After discriminating between knowledge tests and performance tests, he stated, "If [a student] is deficient in the knowledge which is contained in the first test, the chances are good that he will not perform well [on the second test]. If, on the other hand, he answers all the questions correctly, there is still no guarantee that he will [perform] well." To complete Jones's argument, a thorough evaluation of the student's acquisition of the language requires both knowledge and performance tests.

4 The term integration has been used with respect to testing for communicative competence to describe a variety of learning tasks such as the integration of linguistic competence and performance abilities (Carroll 1968:56), that of language skills (Morrow 1979:148), and that of linguistic components and perceptual modalities with language skills (Oller 1980:98). The term is used here with the context of achievement testing to designate a much more fundamental learning task related to Palmer's (1979:170) constructs of "compartmentalized control" and "integrated control".

5 Evans (1980) has reported the development of classroom exercises in university level German language classes which follows a program similar to the one presented here. Her program consists of using the textbook grammar drills in a given lesson to develop the student's linguistic competence and supplementary communicative exercises that refer specifically to the content of the previous lesson to develop his communicative competence.

6 This second procedure is a modified version of the scoring procedures presented by Carroll (1975:78) which have been simplified for use in a classroom testing program.

Bibliography

Bialystok, Ellen. 1979. "The role of conscious strategies in second language proficiency". Canadian Modern Language Review 35: 372-394. Rpt. in The Modern Language Journal 65: 24-35.

Bosco, Frederick J. and Robert J. DiPietro. 1970. "Instructional strategies: their psychological and linguistic bases". Inter-National Review of Applied Linguistics 8: 31-52. Rpt. in Lugton, Robert C. (ed). 1971. Toward a cognitive approach to Second-language acquisition. Philadelphia: The Center for Curriculum Development. 31-52.

Brière, Eugène J. 1980. "Communicative competence, variable rules, and interdisciplinary research". In Scarcella, Robin C. and Stephen D. Krashen (eds). 1980. Research in second language acquisition. Rowley, Massachusetts: Newbury House. 89-94.

Candlin, Christopher N. 1973. "The status of pedagogical grammars". In Corder, S.P. and E. Roulet (eds). 1973. Theoretical linguistic models in applied linguistics. Brussels: AIMAV. 55-64. Rpt. in Brumfit, C.J. and K. Johnson (eds). 1979. The communicative approach to language teaching. Oxford: Oxford University Press. 72-81.

Carroll, J.B. 1968. "The psychology of language testing". In Davies, Alan (ed). 1968. Language testing symposium: a psycholinguistic approach. London: Oxford University Press. 46-69.

Carroll, J.B. 1975. *The teaching of French as a foreign language in eight countries*. New York: John Wiley and Sons. Pp. 47-111 rpt. in Spolsky, Bernard (ed). 1979. *Some major tests*. (Papers in applied linguistics, advances in language testing series, 1). Arlington, Virginia: Center for Applied Linguistics. 1-48.

Chastain, Kenneth. 1969. "The audio-lingual habit theory versus the cognitive code-learning theory: some theoretical considerations". *International Review of Applied Linguistics* 2: 97-106.

Chastain, Kenneth. 1976. *Developing second-language skills: theory to practice*. 2nd. ed. Chicago: Rand McNally.

Cherry, Colin. 1978. *On human communication*. Cambridge, Massachusetts: MIT Press.

Clark, John L.D. 1972. *Foreign-language testing: theory and practice*. (Language and the teacher: a series in applied linguistics, 15). Philadelphia: The Center for Curriculum Development.

Davies, Alan. 1978. "Language testing". *Language Teaching and Linguistics Abstracts* 11: 147-159, 215-230.

Evans, Ilse Klemperer. 1980. "Bridging the gap between pattern drills and communicative exercises". *Foreign Language Annals* 13: 35-38.

Johnson, Keith. 1979. "Communicative approaches and communicative processes". In Brumfit, C.J. and K. Johnson (eds). *The communicative approach to language teaching*. Oxford: Oxford University Press. 192-205.

Jones, Randall L. 1976. "Achieving objectivity in subjective language tests". In Nickel, Gerhard (ed). 1976. *Proceedings of the fourth international congress of applied linguistics*, vol. 3. Stuttgart: HochschulVerlag. 241-253.

Jones, Randall L. 1977. "Testing: a vital connection". In Phillips, June (ed). 1977. *The language connection: from the classroom to the world*. Skokie, Ill.: National Textbook Co.. 237-265.

Jorstad, Helen L. 1974. "Testing as communication". In Jarvis, Gilbert A. (ed). 1974. *The challenge of communication*. Skokie, Illinois: National Textbook Co.. 223-273.

Kleinman, Howard H. 1978. "The strategy of avoidance in adult second language acquisition". In Ritchie, William C. (ed). 1978. *Second language acquisition research*. New York: Academic Press. 157-174

Morrow, Keith. 1979. "Communicative language testing: revolution or evolution?" In Brumfit, C.J. and K. Johnson (eds). 1979. The communicative approach to language teaching. Oxford: Oxford University Press. 143-157.

Oller, John. 1980. "Communicative competence: can it be tested?" In Scarcella, Robin C. and Stephen D. Krashen (eds). 1980. Research in second language acquisition. Rowley, Massachusetts: Newbury House. 95-104.

Palmer, Adrian S. 1979. "Compartmentalized and integrated control: an assessment of some evidence for two kinds of competence and implications for the classroom". Language Learning 29: 169-180.

Paulston, Christina Bratt. 1970. "Structural pattern drills: a classification". Foreign Language Annals 4: 187-193. Rpt. in Joiner, Elizabeth Garner and Patricia Barney Westphal (eds). 1978. Developing communication skills. Rowley, Massachusetts: Newbury House. 21-31.

Rivers, Wilga M. 1975. A practical guide to the teaching of French. New York: Oxford University Press.

Schulz, Renate A. 1977. "Discrete-point versus simulated communication testing in foreign languages". The Modern Language Journal 61: 94-101.

Schulz, Renate A. and Walter H. Bartz. 1975. "Free to communicate". In Jarvis, Gilbert A. (ed). 1975. Perspective: a new freedom. Skokie, Illinois: National Textbook Co.. 47-92.

Tarone, Elaine. 1980. "Communication strategies, foreigner talk, and repair in interlanguage". Language Learning 30: 417-431.

Valette, Rebecca M. 1977. Modern language testing. 2nd. ed. New York: Harcourt Brace Jovanovich.

Zelson, Sidney N.J. 1978. "Skill-using, self-expression and communication: exercises in three dimensions". In Joiner, Elizabeth Garner and Patricia Barney Westphal (eds). 1978. Developing communication skills. Rowley, Massachusetts: Newbury House. 44-56.

ED217730

REPORT ON AN EXPERIMENT IN ACTIVATED GROUP ACHIEVEMENT TESTING FOR ADVANCED STUDENTS OF SPOKEN ENGLISH

T. Hayward, A Michiels, and J. Noël (University of Liège)

This experiment is one of a series attempting to overcome practical constraints on testers (e.g. limited time in the curriculum, shortages of staff and the convenient comforting availability of machine test marking,) which frequently make it much easier to test grammatical competence rather than communicative performance in spoken English. Although large numbers of students can now be processed very quickly with minimal manpower, examiners have become test writers and statisticians rather than being encouraged to make decisions on ability.

This paper examines an experimental test based on rehearsed communicative performance through activation which tries to approximate to authentic communicative behaviour in spoken English, and, in addition, presents an evaluation scheme currently used for assessing group and individual performance.

The subjects of the experiment were two hundred and fifty students in their first and second years of study for a joint language degree who were following a twelve-week spoken English course based largely on group work and guided self teaching. Books in use were Functions of English (Jones 1977) and Well Said (Dixey and Rinvolucri 1974). The test was used in the course to assess achievement both during and after the course.

Following each "learning phase" which was of either two or three weeks duration, each group of four students was presented with a written description of an incident or scene (see Appendix 1 for an example). (1)

This was designed to provide social roles, outline character, settings and topics linked to the language functions taught in the previous weeks and which were to be tested. Each group of students was required to write in dialogue form a script and produce this as a cassette recording of between five and ten minutes length, which was then evaluated. For the final examination no description was given, leaving students with a free hand as to their own ideas, excepting only that they had to use a significant number of functions from their course. (Appendix 2 shows a sample from a script). Preparation time was varied from two days to two weeks.

Both cassette recordings and live performance were assessed using a mark scheme consisting of four elements (2) : Authenticity, Interaction, Idiom/Register and Intonation which together cover manageably the range of skills which the test required. The scheme was made available and explained to the students involved the experiment (Appendix 3 shows the mark scheme).

FL013020

Authenticity can be defined as the quality of the social role, setting and characters as shown in the dialogue and any other effects (e.g. traffic noise introduced in the background to simulate a street scene). The levels in this element of the mark scheme attempt to reflect the imagination, effort and effect of the group production. It is clearly most difficult to score 1.0, but the maximum has been achieved on several occasions. In practice, a score of 3.0 is a good average mark, with about 4.5 the usual top level. Authenticity can broadly be summed up by the examiner asking the question "How far does this approximate to a real scene?"

The second element, Interaction, assesses the speakers' actions, and reactions to each other and the setting through the language, without requiring a competence in drama from the student. In other words, the students' behaviour is compared with that of a spontaneous native speaker, through examination of such things as delivery speed, interjections, tags, spontaneity etc., during the dialogue. Two points should be noticed here: communicative effectiveness and not grammaticality is the concern; a group performance has to be some way below average before the communication is seriously hindered or destroyed, implying that a certain minimum level is expected from the students. Although we talk of group assessment here, individuals can be assessed because differences in Interaction can be particularly marked, whereas establishing Authenticity is impossible.

For the third element of the mark scheme, we found it better to group Idiom and Register, for practical reasons. The scale ranges from naturally idiomatic language in a contextually appropriate register, to translations and grammatically inaccurate sentences which have minimal communicative value. Although not exclusively, poor results in this element will manifest themselves in slavish and inappropriate borrowings from the coursebooks and forced uses of grammatical or functional stereotypes found in some of the so-called situational materials, especially those published in the fifties and sixties, but which are sometimes still in use.

Intonation, the fourth element of the work scheme, may seem an obvious choice in a test of spoken English, but because of the variations between members of a group, it poses problems for assessment. It is for this reason that the description has to be as broad as possible, ranging from native speaker level to a source language intonation dominating to the point that communication is destroyed. Poor performance by our (largely) native speakers of French may be assigned to their difficulty in producing genuine English stress and rhythm: for instance, no systematic use of weak forms, appropriate attitude markers and interaction signals. We have highlighted word stress in particular because we feel that it is an area which is genuinely reflective of a student's general intonational ability and one where hard work can make a significant difference to individual and group performance.

In use, with cassettes and a single assessor, listening evaluation is carried out using a sampling technique, each element being taken in turn for a total of about sixty seconds. Once a rough level has been established, the remaining tape time is used for refining, revising and for making quick correction notes on the script, for which a series of indicative symbols is being developed. By being able to assess during the running time of the tape, speed of processing is increased, so that up to ten cassettes per hour can be dealt with.

For panel decisions on live performance, each member gives a personal assessment using the mark scheme and the mark is agreed upon by quick consultation at the end. Both the group and the individual are assessed in this section although the Authenticity element is clearly ignored for the latter. During an eight minute presentation, the examiners have to make twelve decisions concerning individuals and four for the group, something that, along with Falk and Strange (as outlined in paper delivered to the I.U.S. conference, Brussels,1980) we found no difficulty in doing. The same holds for obtaining agreement amongst panel members.

In both types of performance the group and the individual are given a mark out of twenty and in addition short comments indicating the problems or strengths. Both final examination and cassette work are counted for grade purposes but given equal weighting.

The attraction of this test where there are large numbers of students is that they are involved in interactive exercises which require the linking of competence and performance and the presentation of these through activation. Skills are reinforced and practised until a satisfactory level is achieved by the group. By the eventual rapprochement of student awareness and testers' expectations, achieved by sample tapes, a heightened level of preparation and out-of-class participation is ensured, benefitting test and course work. Similarly, the inter-reliance of individuals in the group fosters a critically constructive faculty which is most valuable in group language learning.

The preparation process is of particular interest and recordings of this with different groups of students reveal that a very conscientious amount of rehearsal and coaching takes place, besides enabling the student to prepare at his own convenience. An additional advantage is that precious class time is not impinged upon.

Student groups become involved in the subject matter they are given and the response of their imaginations is particularly rewarding, their enthusiasm running to 45 minute productions and costume performances. Most tests and a surprising number of course-books fail, in our opinion, to exploit this aspect of the learner's capacity.

Granted that some groups felt that the preparation period was too long, this time is a very valuable link between the test and the course and is heavily used by the students to perfect themselves, to research and to help each other, an aspect we try to foster by insisting that the whole group is responsible for each individual's performance. This type of test is then best compared with group seminar paper or seminar work.

A particularly attractive feature of the experiment is the involvement of the tester in the making of rapid assessments of communicative performance based on experience and established criteria, in contrast to interpreting statistics. Although there was some hesitation initially, evaluations were made with increasing self-confidence with the concentration on relatively minor errors being gradually replaced by a willingness to sample mark and judge the overall effect of the work. (We are at present undertaking a more detailed study of ways of achieving long term consistences amongst individual assessors working alone). In the light of the exciting developments taking place in communicative teaching methodology and materials (e.g. *Encounters*, by Garston-Sprenger, Jupp, Milne, and Prowse 1979, and *The communicative teaching of English*, Candlin 1981), it seems appropriate that the role of tester should be reassessed, especially if such elements as Interaction are to be evaluated accurately. It is tempting to go further and suggest that the tradition of individual assessment within language teaching should be modified to include some element of the group performance, but the problems inherent in this approach are legion. However, it does seem sound to suggest that where, for instance, a traditional interview format is used to test spoken English ability, then a group assessment should be considered alongside.

Finally, it is felt that the use of a test format which resembles a class exercise brings teacher and tester closer, together and ensures that evaluation retains a relevant position in the syllabus, not dominating and twisting it, as has happened with the TOEFL examination, for example.

Bibliography

Brumfit, C. J. and K. Johnson (eds). 1979. *The communicative approach to language teaching*, Oxford: OUP.
Candlin, C. N. (ed). 1981 *The communicative teaching of English*. Harlow: Longman.
Dixey, J. and Rinvolucri, 1974. *Well said*. Canterbury: Pilgr
Jones, L. 1977. *Functions of English*. Cambridge: CUP.
Widdowson H. G. 1978. *Teaching language as communication*. Oxford: OUP.

Garston-Sprenger, J., Jupp, T. C., Milne, J., Prowse. P., 1979. Encounters. London : Heinemann.

Appendix 1

Our scene this time takes place on the road border between Lumbago and the small state of Liechtenbourg. It is the habit of the citizens of Lumbago to deposit money in the small neighbouring state in order to avoid paying taxes and also to buy cheap goods which they smuggle through the indulgent customs. However, under the direction of their tough new chief, Major Quiixhote Andreus de Moulin, the customs have been cracking down on the Lumbelgians, but they are going out of their way to be polite and avoid ugly scenes, especially as the representative of the local press, a virulent and caustic old scribbler Pepe Fouetard is there looking for trouble. However, trouble there is because the good citizens of Lumbago do not willingly take to being stopped and searched.

A car approaches the border post and is flagged down by the polite customs officer. He asks the occupants whether they have anything to declare and on hearing that they have nothing, he asks the car to pull over to be examined. The driver, a young spoilt student, immediately begins to bully and bluster, reminding the officer that his father is the Mayor or Nerviers. Politely, they are made to open the boot of the car where a considerable amount of contraband is hidden. The other passengers of the car, Sylvia Allemand, who is shy and timid and Fulvia Yatterback, an empty headed model, who is after the driver's money and body, enter the arguement. You should develop the dispute which is brought to a conclusion by the arrival of Major Moulin and a gendarme, and by a hefty fine levied on the car's occupants. The scribbler constantly tries to provoke the participants and breaks into the conversation with exaggerated remarks which the others are forced to correct. Elements of the dispute should include (amongst others) attempts to bully the officer, a subtle attempt to bribe him, a statement of total disbelief, an attempt to throw themselves on his mercy, the 'I agree officer but can't you let us go this once approach' and the grumbling acceptance of his decision. Their attitude to Moulin should be a total contrast as should that of the customs officer.

Good luck

Appendix 2

COFFEE MORNING

SHIRLEY - (pouring out tea) Sugar in your tea ? ... Fiona ?
FIONA - Yes, please.
DORIS - I wouldn't in your place !
(to Shirley) I don't have any, dear, ...
SHIRLEY - (handing a cup of tea to Doris). Here you are, Doris... This coffee morning is quite a success... Look at all the people... Isn't that Ann Theak, stuffing herself with flapjacks ?
FIONA - Yes; have you heard she's engaged ?
DORIS - Engaged ! I hope she will jump at the chance ! Isn't she already 42 ? ... And she's so wobbly, ... and spotty ... and grumpy ... and her hair is greasy ... Who's the lucky man ?
SHIRLEY - Mr. Bloodyknife, the butcher. They met at the stock-market; she was seduced by the strength of the muscles, the sweet glance of the eye ... of Bacchus, his favourite bull.

LYNN comes towards the 3 ladies sipping their tea

LYNN - Hello ! Having a good time ?
SHIRLEY - Yes, why don't you take a chair and join us ? ... I am Shirley, this is Doris ... and Fiona.
LYNN - I'm Lynn. I don't know many people yet but isn't that Mrs Chatterton over there ?
FIONA - Yes, she's talking to Miss Withebooks, the new librarian ... (looking around). And who's that in the white dress ? I remember her face but her name eludes me at the moment.
DORIS - It's Purity Devill.
LYNN - Beautiful day for a coffee morning, isn't it ?
FIONA - Mmm ... I'm glad I could make it ... Actually I should be at a meeting of the N.A.P.B.K. but I was so anxious to get acquainted with our new neighbour, Mrs Watney.
LYNN - You said the N.A.P.B.K. ? What's that ?
FIONA - National Association for the Protection of Battered Kittens; I just volunteered to be their deputy chairman ...
I've been so busy since George died ... you know, being a member of the Widows Front, and the National Aid to the Bald ... They've also asked me to help with the raffle at the hospital fête this year.
DORIS - (ironically) Oh they have, have they ? Indeed Fiona, you seem to have very busy days ... You must meet some interesting people ... Perhaps one day you'll marry the president of the committee for the help of pink frogs !!
FIONA - Doris !
LYNN - (trying to smoothe things down) ... Er ... Would you like some shortbread ? By the way, have you been to Modesty ? They've got some nice jeans and very cheap indeed.

DORIS - Lynn, do you really like jeans ?

LYNN - Well, they're very casual ...

SHIRLEY - Look over there ! Isn't that M. Holycross, the new vicar ? I quite like him ... he's young ... and handsome ! The children like him ...Ian appreciates him wearing a silver cross.

DORIS - (contemptuously) That's icing on the cake, dear !

LYNN - What's wrong with that ?

FIONA - Anyway, he's not just a vicar, he's the one who's been attending the most coffee mornings since he arrived in the parish !

DORIS - Oh, look at your son, Shirley ! He's pulling the tail of the Grumble's dog !!

SHIRLEY - Christopher ! Christopher !! Leave that dog alone, will you ! Otherwise I'll tell your daddy about it tonight ! ... And where's your sister ? ... (looking around) ... Oh Sandra ! Do put that mug back on the table ! I would hate you to drop it !

LYNN - Oh, you have a lovely skirt, Fiona. Where does it come from ?

FIONA - Thank you ! Well, it comes from ... Liberty's.

DORIS - (mumbling) Oh it does, does it ? I saw exactly the same for 10 pounds in Marks & Spencer last week ! ...

LYNN - (embarassed) Some more tea, anybody ? ...

DORIS, FIONA, SHIRLEY - No thanks !

LYNN - Whose are those twins, digging out the crocusses at the bottom of the garden ?

SHIRLEY - They're mine.

LYNN - Shirley, ... how many children do you have ??

SHIRLEY - Six.... At the moment that is ...

DORIS - Oh, I do hate children ! Take the Breakalls for instance, always skating down the estate. The other day, I was peeping through the curtains as usual and do you know what little John did ?? He waved at me !!!... Those children !

Appendix 3

Evaluation Scheme for Assessing Spoken Performance in Cassette and Live Presentations.

1. Authenticity

Mark	Level Content
5. 0	Sounds like native speakers talking in an authentic setting. Exhibits complex attitudes easily and mixes them in related contexts. Consistent and clearly developed characters reflected in the language. Highly imaginative content, brilliant context demanding attention.
4. 0	Clear setting. Attitudes used are somewhat simple and characters portrayed are consistent with some reflection in the language. Good,bright, well thought out content.

Mark	Level Content
3. 0	Some attempt at setting. Attitudes shown are not clearly reflected in the language. Any characters portrayed are not consistent. Appropriate content showing thought and imagination.
2. 0	No setting. No consistent attitudes. Participants fail to adapt themselves to roles. Some attempt at characterisation and student involvement, but minimal. Unimaginative content.
1. 0	No setting, no attitudes, no characterisation. No student involvement is apparent. Content is insufficient and sounds unplanned.

2. Interaction

5. 0	Speakers interact with each other in a natural way. Dialogues include authentic interjections, tags, interruptions etc. Free and authentic use of attitudes which are stimulated from other speakers. Variable speed to match content.
4. 0	Speakers listen to each other and react almost naturally. Some interjections and a few interruptions almost accurately used. Attitudes and opinions have a high degree of spontaneity. Delivery speed matches content.
3. 0	Speakers are reacting from a written script and the dialogue has a synthetic quality. Replies are carefully balanced with no interjections etc. May be inappropriate. Delivery speed is similar to native speakers, but not varied to suit content.
2. 0	Delivery speed slow. Referring to a script. No spontaneous reactions to what is said. A series of monologues loosely related. Errors seriously hinder communicative ability or input.
1. 0	Reading a script very slowly with long pauses. No reactions to what is said. A series of unrelated monologues. Errors destroy input.

3. Idiom/Register

5. 0	Free use of a wide range of idioms appropriate to the dialogue setting and characters involved. Registers changed to suit variations. Some use of slang and abbreviated forms.
4. 0	Wide use of idiomatic English including phrasal verbs and 'got'. Good consciousness of idioms and registers appropriate to the situation, though not perfect. Some use of shortened forms.
3. 0	Limited use of idiom mixed with grammatically correct language though not direct translations (or faux amis). Some abbreviated forms. Registers mixed and not well controlled. Occasional failure to use suitable word.

Mark	Level Content
2. 0	Statements are generally grammatically accurate but with translations and faux amis. Little accurate idiomatic usage. Sounds like a grammar book talking. Language also restricted to typical 'school' English. No appropriate register.
1. 0	Grammatically inaccurate sentences without noticeable idiomatic content. Translations from source language with many faux amis. No consciousness of register.

4. Intonation

5. 0	Almost sounds like a native speaker talking
4. 0	Intonation used for statements, questions, tags and interjections is almost accurate. Still some mistakes when handling moods/emotions and attitudes. Ability to use stress appropriate to context. No misstressed words.
3. 0	Limited success in intoning the elements of 4. 0 but some exaggeration. The correct intonation may sound scripted and deliberate. Mixed source/target language patterns. Few misstressed words.
2. 0	A few attempts to intone, but largely source language dominated. Communication seriously hindered by source language. Noticeable number of wrongly stressed words.
1. 0	Flat, unemotional monotones or totally source language, showing no consciousness of target language intonation at all. Communication destroyed by source language interference. Very high number of wrongly stressed words.

NOTES.

1.- The slightly tongue-in-cheek nature of some of the written descriptions was deliberate, in order to try to evaluate which produced a better student response and participation. It seems that the more imaginative the better.

2.- The mark scheme was developed from the authors' experience in evaluating group performance in E.F.L. both recorded and live, especially in E.S.P., where similar problems were posed by groups of students carrying out scientific experiments in laboratories, using English as their medium. It should be noticed that the elements fall into two groups : those which can be researched, practised and "learned" at the individual level (Idiom/Register and Intonation), and those which require the development of skills for group interaction and which can only be developed through communication (Authenticity and Interaction) and which come from accumulated experience. From listening (unobserved) to groups preparing for the test, it was clear that the elements were interpreted in this way by students too.

The isolation of the skills involved in the group test largely came about through observing and listening to group interactions of various kinds. It must be recognised that the elements are tightly interrelated and that their separation is largely pragmatic. Authenticity can involve a greater cultural awareness than is suggested by the definition although we have attempted to minimize this through careful selection of test descriptions.

Experience would also seem to suggest that the ability to use certain language features adequately is indicative of levels within the elements. In Interaction, for example, we have highlighted 'tag' endings because it is felt that at level 4.0 and above, they should be present with a certain degree of freedom. Similarly, Idiom/Register cites "got" at level 4.0 for the same reasons.

ED217731

THE DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER ACHIEVEMENT TEST PROGRAM

Robert G. Henderson (Monterey)

Some of you may be familiar with the Defense Language Institute Foreign Language Center and some of its programs. Over the years we have been host to a steady flow of visitors from Europe, Asia, Latin America and Canada. For those who are not knowledgeable about our school, allow me to supply a few highlights.

DLI is the proponent agency for the Department of Defense Foreign Language Program. Our primary responsibility is to provide language training for all four military services, as well as for a few other federal agencies such as the Federal Bureau of Investigation and the Central Intelligence Agency. We currently train over 4,000 men and women each year in some twenty-eight languages. At any particular point in time we have approximately 2,500 students on campus at the Presidio of Monterey, California. A faculty and staff of over 800 is required to operate the school and manage other functions for which DLI is responsible. We also maintain liaison offices in Washington D.C. and at Heidelberg.

Our job/task analysis specialists have identified over three-hundred military occupational specialties that require the use of a foreign language in one way or another. It would be, of course, astronomically expensive to give each student the specific language skills necessary for a particular job. In reality, the bulk of our students are represented by about half a dozen military jobs.

The institute is organizationally structured into five directorates. The Director of Training is responsible for both resident and non-resident training. The Director of Evaluation is our quality assurance group. The Director of Information and Resource Management houses our financial, manpower and accounting functions as well as our computer and automatic data processing operations. The Director of Support is our housekeeper.

The views expressed by the author do not purport to reflect the position of the United States Government, the Department of Defense, the United States Army, or the Defense Language Institute.

FL013021

Major course and test development projects are the responsibilities of the Directorate of Training Development. These projects are carried out through a team approach in which each member brings a special area of expertise to the work. Job and task analysts, curriculum design specialists, test specialists and course writers who are native speakers of the Target Language all contribute to the process.

The Test Division is responsible for aptitude, achievement and proficiency testing in the Defense Foreign Language Program. Although in psychometric circles these terms are still somewhat controversial, for our own purposes we have established working definitions as follows.

APTITUDE TEST - A test designed to give a quantitative estimate of a subject's ability to profit by training. (Chaplin 1968)

ACHIEVEMENT TEST - A test designed to measure the effects of a specific program of instruction or training. (Anastasi 1968)

PROFICIENCY TEST - A test designed to measure the subject matter competency of an individual regardless of the means by which he or she obtained the knowledge, skill, or ability. (Henderson 1974)

Our aptitude test program is represented by the Defense Language Aptitude Battery. This is a 90 minute pencil and paper test which includes an audio component. It is used primarily at Armed Forces Entrance and Examining Stations and Induction Centers to identify potential foreign language training candidates. It was introduced in 1977, replacing the Defense Language Aptitude Test, which had been in use for 25 years.

Our proficiency test program is represented by the Defense Language Proficiency Test series. These are pencil and paper tests with an audio component which measure listening and reading comprehension in 44 languages or dialects. The tests are used both for our students at the time of graduation and on a worldwide basis for all military personnel who claim a compentency in a foreign language. The system is being continuously upgraded with the addition of new tests and alternate forms of existing tests.

Our achievement test program must respond to a variety of resident and non-resident courses of instruction. Among its responsibilities, DLIFLC is also responsible for supporting command-sponsored language training. Currently we are engaged in developing course and test materials for four types of non-resident courses.

1 HEADSTART

Headstart is a modest introductory program designed to provide any interested individual with enough language familiarization and cultural awareness to better appreciate and enjoy his or her stay in the host country. The Test Division supplies the end-of-course examinations.

2 GATEWAY

Gateway is also an introductory program especially designed to provide senior military commanders with an elementary language capability for special situations. It enables our personnel to understand and participate in such activities as change of command ceremonies, lunch with the mayor, social and work-related discussions with military peers, etc. These programs are especially helpful for senior officers assigned to US Army Europe, as commanders of units stationed in Germany, and those working in the Combined Forces Command in Korea. The Test Division supplies the end-of-course examinations.

3 REFRESHER MAINTENANCE

The Refresher Maintenance courses are designed for use by individuals who are, or once were, using a foreign language as part of their military duties. The major purpose is to maintain proficiency, to combat language skill loss, or to regain lost competency. The course design is based primarily on the last phase of the resident Basic Course. Except for technical consultation, the Test Division is required to give only minimum support since the tests used are drawn directly from the Basic Course measurement programs.

4 TRAINING EXTENSION

Training Extension Courses are somewhat more ambitious. They are designed not only to maintain existing language skills, but to augment and enhance language ability in technical specialties. At present our work is limited to programs for the interrogator field. The Test Division provides only module tests for this program, since end-of-course tests are not required.

The non-resident program is not our largest consumer of achievement test development. Our resident courses command that distinction.

Since 1970, and more vigorously since 1977, DLI has been struggling with a multifaceted problem regarding the resident training continuum required to meet the needs of widely differing military jobs. Quite recently, the shape and scope of the training continuum has become more clearly defined as we learned how to cluster similar needs and our users also articulated their requirements better in terms of language application objectives. The fundamental problem is that we have difficulty in retaining our students long enough in resident training to learn the quantity and quality of language that is essential for their jobs. The basic challenge is to provide adequate training, yet leave sufficient utilization time on the job before the individual's term of enlistment is over. Despite reasonably well-tuned instructional programs, bright, motivated students and an intensive study environment, we are not able to train students to the level of competencies needed in as short a time as six months for languages such as Italian and Spanish or twelve months for Chinese and Russian. DLIFLC and the military personnel systems of the four services are continuing to grapple with this problem.

In general, future resident curricula envisage a series of phases of instruction. The first three phases would approximate the instruction now reflected by our basic courses. Phases four and five would reflect the instruction now allocated to intermediate and advanced courses, plus whatever specialized instruction may be necessary to respond to requirements identified from job/task analysis, user agency stipulations, or other sources.

In September 1979 DLI published a Basic Course Design document. This document stipulates that our resident courses will include module tests, qualifying tests and final examinations.

BASIC COURSE TEST REQUIREMENTS

Instrument*	Skill Tested	Format
Module Test	Various	Special
Qualifying Test	Listening Comprehension	Multiple Choice
	Reading Comprehension	Multiple Choice
	Speaking	Taped Interview
Final Exam	Listening Comprehension	Multiple Choice
	Reading Comprehension	Multiple Choice
	Speaking	Taped Interview

---

* All tests prepared in minimum of two alternate forms

Module tests must faithfully reflect the content of the learning module itself. Therefore, each module test design is unique while incorporating standardized test formats and scoring systems where feasible. The test specifications are derived from the learning objectives stated for each module. The test specifications are stated in the familiar behavioral terms of stimulus, task, conditions and criterion. Here is an example of the test specifications for the Written Interpretation exercise for Module #1 of the Korean Basic Course now under development. Three additional skills for Module #1 are treated in the same manner. All module tests are prepared in a minimum of two alternate forms.

KOREAN BASIC COURSE

Module #1 Test Specifications for Written Interpretation

1 Stimulus - Ten sentences recorded in Korean.

2 Task - Translate each sentence into equivalent written English.

3 Conditions -

a. All vocabulary and grammatical features expressed in Korean will be limited to the vocabulary and grammar presented in Module 1 (Lessons 1-5) of the Korean Basic Course.

b. Each recorded sentence will be voiced twice in immediate succession with a three-second pause between repetitions.

c. Examinee response time is automatically controlled by the tape.

d. The use of lexical aids is not permitted.

4 Criterion - The minimum acceptable performance standard is 75%.

Qualifying Tests and Final Examinations can be fully standardized (with some minor variations to accommodate unique orthographic systems). The Qualifying Tests are used for making important personnel decisions concerning retention in training. First, the student must successfully "pass" all preceding Module Tests in order to be permitted to take the Qualifying Test. The test is administered at about the tenth week of instruction for shorter courses (French or Portuguese, for example) and about the twelfth week for longer courses (such as Japanese and Polish). The minimum acceptable performance standard (often called the cutting-score) for the Qualifying Test is arbitrarily and tentatively set at 75 per cent. Changing attrition rates or other factors could influence the school to change the cutting-score. Second, if the student surpasses the cutting-score he or she proceeds to the next phase of instruction. If the student's score on any of the three tests in the battery falls between 70 and 74, the student is placed in probationary status. Counseling is conducted by the student's training advisor (who is usually a language qualified officer), since poor test performance may be due to reasons other than academic (attitude, motivation, health, anxiety, etc.). If the problem appears to be an academic one, remedial instruction is then arranged and the student is subsequently retested. If a second failure occurs and is judged to be academic in nature, the student may be either recycled (to a following class or a different language) or disenrolled. The training advisor has a considerable amount of information on which to base this judgment.

When a student scores below 69 per cent on the Qualifying Test he or she is again referred to a counselor. The student may be recycled or summarily dropped.

The central purpose of the Qualifying Test, taken together with other information, is to identify potential academic losses as early and accurately as possible. Current academic attrition runs at about 13 per cent (and is creeping up slowly). Of this group, the majority are concentrated in relatively difficult languages (such as Arabic or Chinese) and almost all occur prior to the twelfth week of instruction. With current training costs averaging $170 per week per student, each day of delay in identification of these possible losses becomes important to cost-effective training operations. It must be remembered that, except for Module Tests, the student will not again encounter a major test battery until the end of the course.

Now let's look at the test matrices for the Qualifying Test.

First we have Listening Comprehension. (See next page). Note that, with the exception for the Situational Response section, all items are bilingual. And, with the exception of Situational Response, all four options are printed in the test booklet.

The Situational Response item stimulus is recorded on tape, just as the other items are. However, the four responses are also taped in the Target Language. Though this places a considerable load on examinee memory, students do respond to it rather well. The item takes the form of a mini-dialog. The examinee's task is to select the appropriate second or fourth line of dialog from among the four options presented. The other sub-sections on the Listening Comprehension are quite conventional.

BASIC COURSE QUALIFYING TEST
Listening Comprehension Matrix

Vocabulary Identification

Stimulus - Target Language
Options - English

Difficulty Index	Item Number
Easy	1,2,3
Medium	4,5,6,7,8,9
Hard	10,11,12
Sub Total	12

Translation

Stimulus - Target Language
Options - English

Difficulty Index	Item Number
Easy	13,14,15
Medium	16,17,18,19,20,21
Hard	22,23,24
Sub Total	12

Situational Response

Stimulus - Target Language
Options - Target Language
(voiced only - no print)

Difficulty Index	Item Number
Easy	25,26,27,28
Medium	29,30,31,32,33
Hard	34,35,36,37
Sub Total	13

General Comprehension

Stimulus - Target Language
Options - English

Difficulty Index	Item Number
Easy	38,39,40,41
Medium	42,43,44,45,46
Hard	47,48,49,50
Sub Total	13

TOTAL	50

A test length of fifty items is based on econometrics. We purposely overwrite the test by ten items. The item data acquired during the field trial process are then analyzed to select the forty items with the best discrimination and difficulty indices. We expect the reliability coefficient on the final version to be about .78, using the Kuder-Richardson Formula 20. Our self-imposed standard is a KR-20 of .75. This is quite important, since each increase in test length of ten items carries an estimated production cost of $730. Test administration time is also an important aspect of test length because every minute a student spends in testing becomes a minute subtracted from instruction.

The next graphic shows the test matrix for the Reading Comprehension portion of the Qualifying Test. (See next page).

Again, Vocabulary Identification is almost the same as its Listening Comprehension counterpart. The single difference is that the stimuli, as well as the options, are printed in the test booklet. The same is true of the Target Language to English translation section. The third section is unique to the Reading portion. In this case the stimulus is printed in English and the options are printed in the Target Language. The need for this skill, which we used to neglect in earlier formats, was uncovered as the result of job/task analysis. It is, of course, much less emphasized than Target Language to English translation.

Section Four of the Reading Test is also a newly-introduced item type. It takes the form of a modified cloze technique adapted to a multiple-choice format. A printed Target Language sentence of approximately fifteen words has three blank spaces inserted in it, each space labeled with a number. Below the sentence three corresponding numbers are printed, each accompanied by four options. The examinee's task is to select the word, sometimes part of a word, that if inserted in the blank space would make the Target Language expression grammatically and syntactically correct.

The last section, General Comprehension, differs from its Listening counterpart by being monolingual. Both stimuli and options are printed in the Target Language in the test booklet. We are sometimes asked why we do not format the Listening General Comprehension items in the monolingual arrangement. The reason is that validity ("What is being measured?") becomes a question. If the Target Language options are printed in the test booklet we force the examinee to read in that language. The examinee may clearly hear and understand the taped stimulus, but still fail the item due to poor reading comprehension in the Target Language.

BASIC COURSE QUALIFYING TEST
Reading Comprehension Matrix

Vocabulary Identification
Stimulus - Target Language
Options - English

Difficulty Index	Item Number
Easy	1,2,3
Medium	4,5,6,7
Hard	8,9,10
Sub Total	10

Translation
Stimulus - Target Language
Options - English

Difficulty Index	Item Number
Easy	11,12,13
Medium	14,15,16,17
Hard	18,19,20
Sub Total	10

Translation
Stimulus - English
Options - Target Language

Difficulty Index	Item Number
Easy	21,22,23
Medium	24,25,26,27
Hard	28,29,30
Sub Total	10

Contextual Comprehension
Stimulus - Target Language
Options - Target Language

Difficulty Index	Item Number
Easy	31,32,33
Medium	34,35,36,37
Hard	38,39,40
Sub Total	10

General Comprehension
Stimulus - Target Language
Options - Target Language

Difficulty Index	Item Number
Easy	41,42,43
Medium	44,45,46,47
Hard	48,49,50
Sub Total	10

TOTAL	50

The Speaking Test is a significant departure from our former practices. For years DLI has relied heavily upon face-to-face oral interview technique developed by the Foreign Service Institute, which is the language training school for our State Department. Although a very effective methodology, the format requires two native, or near-native, examiners for each examinee and twenty to thirty minutes to administer and score. Beginning in 1976 we were able to modify the format to use only a single examiner in a time span of less than twenty-five minutes with no loss in fidelity. The technique was described in detail by Dr. Francis Cartier in "Measuring Spoken Language Proficiency" in 1980.(1) We are very conscious of the importance of instructor contact hours in our resident courses and try to minimize the consumption of instructor contact time for testing without degrading the effectiveness of our tests. When graduating about 4,000 students annually it can be seen that the resource consumption numbers are considerable.

In the mid-1970s we began experimenting with a taped oral production measurement technique. We had two objectives in mind. The first was to find a way to make an oral production test exportable and associate it with our worldwide Defense Language Proficiency Test program, which now measures only listening and reading. The second objective was to find a way to administer a speaking test to groups of examinees, as opposed to individually under the existing system.

The first objective, exportability, was partially solved. The need for knowledge of the Target Language to administer the test was eliminated. A test manual is provided for an English speaking administrator. We could not solve the requirement for native language competency for scoring the test in the field. This, of course, is no major problem on the DLI campus. For field use we recommend sending the examinee's tape to DLI, FSI, or to any agency where trained raters are available and where a prior arrangement has been made to accomplish scoring of the test. This has the unfortunate requirement of delaying the use of test results for administrative purposes. We do permit field scoring of the test if qualified personnel are available.

Notes

1 Measuring Spoken Language Proficiency. James R. Frith, ed. Georgetown University, Washington D.C. 1980, pp.7-14.

The second objective, group administration, was solved. This does require use of a Level III language laboratory with recording capability at each position. We are fully equipped with these laboratories at DLI. Use of the test in the field rarely requires group administration. Individual testing is easily conducted using regular reel-to-reel or cassette equipment.

While the taped test requires more time to administer and score than the face-to-face interview, other benefits seem to outweigh these deficiencies from our point of view. One benefit is derived from dramatic savings through group administration. At DLI we can give the test to eighty students at a single sitting. Another benefit is standardized administration and scoring procedures. This tends to reduce the subjectivity inherent in the face-to-face interview. Since the responses are captured on tape, we can use as many independent scorers as we like. Interscorer reliability on the two forms we currently use in Spanish is .95. Correlation between the taped test and the face-to-face interview is .73. And, as noted earlier, the test is exportable. Correlation of test score to course grade, one measure of validity, is .87.

With that rather prolix introduction, let's take a look at the Qualifying Speaking Test matrix. (See next page).

The test is divided into four sub-sections. Given the test length you see here, about thirty minutes are required to administer the test. The first section, Vocabulary Identification, presents a series of thirty-four cartoons or line drawings depicting an object or an action. The taped voice asks, in the Target Language, "What is this?", or, "What is the person doing?" At a tone signal the examinee voices his or her response, which is recorded on a separate tape. The stimuli are arranged in order from easy to difficult. Their selection is based upon word frequency lists, course materials and terms used in face-to-face interviews by individuals at different points in the training continuum. Though a straightforward and relatively simple exercise, the Vocabulary Identification section has the purpose and effect of increasing test score range.

The second section of the test, Structure, presents a series of thirty printed English sentences accompanied by an incomplete, printed Target Language translation of the English sentence. At the sound of the tone the examinee voices the full Target Language sentence, supplying the missing component. The structural elements are again arranged from easy to difficult using the same technique applied to Vocabulary Identification. This section is very useful for determining the extent of grammar control.

BASIC COURSE QUALIFYING TEST
Speaking Matrix

Vocabulary Identification

Stimulus - Printed picture or cartoon with voiced TL question.

Response - Voiced TL description of pictured item or situation.

Number of Items - 34

Structure

Stimulus - Complete printed English sentence. Incomplete printed translation of English sentence.

Response - Voiced complete TL sentence.

Number of Items - 30

Composite Communication

Stimulus - Situation printed and voiced in English.

Response - Voiced TL role based on situation.

Number of Situations - 4

Directed Response

Stimulus - Short story printed in English. TL question printed and voiced.

Number of Stories - 2

Number of Questions - 10

The third section, Composite Communication, may be the most important part of the test. The examinee is presented with four everyday social or business situations printed in English and voiced in English on the tape. The examinee is given a few moments to compose a short speech, playing a role designated by the instructions. At the signal the impromptu speech is voiced. Points for loquacity and poetic expression are minimized by the scoring system since it is not composition ability that we wish to measure, but simply adequate control of the language to deliver a message or request. The role-playing situations are presented, and responded to, one at a time.

The final section, Directed Response, presents two short stories printed in English in the test booklet. The examinee is allowed sufficient time to read the story. A series of five questions, based on the content of the story, are printed in the Target Language in the test booklet. The same questions, one at a time, are voiced on the tape. Following each question the examinee hears a signal and voices a response to the question in the Target Language.

The rater used a special scoring sheet with separate areas for each of the four sub-sections of the test plus "bonus" points for pronunciation. The scoring system is weighted to reflect what we believe to be the more important subskills in oral production as measured by this test. The total possible score on the Qualifying Test for Speaking is 85 points. Approximately thirty minutes is required to administer the test and eight to ten minutes to score it.

The Listening, Reading and Speaking Qualifying Tests are designed to be administered and scored independently and in no specified order. We do insist that they be administered rather closely together in time and that intertest scores not be averaged. The criterion cutting scores have been tentatively established at 75%. Language Department Chairpersons may raise this score at their discretion, but may not lower it. The fact of the matter is that our chairpersons have such well-tuned experience that they can identify probable student academic losses as early as the sixth week with astonishing accuracy. The Qualifying Test is, therefore, a largely administrative tool used to confirm predictions already made intuitively.

We can now look briefly at the Final Examinations matrices.

You can see that the Listening Comprehension matrix differs only slightly from the Qualifying Test. (See next page). The Vocabulary Identification section has been dropped. Test length has been extended to seventy items. Once again we overwrite by ten items and, using item analysis, select those items with the best discrimination and difficulty indices. The final, sixty-item version of the test will yield a theoretical reliability coefficient of .88, using KR-20.

BASIC COURSE FINAL EXAMINATION
Listening Comprehension Matrix

Translation

Stimulus - Target Language
Options - English

Difficulty Index	Item Number
Easy	1,2,3,4,5
Medium	6,7,8,9,10,11,12
Hard	13,14,15,16,17
Sub Total	17

Situational Response

Stimulus - Target Language
Options - Target Language
(voiced only - no print)

Difficulty Index	Item Number
Easy	18,19,20,21,22,23,24
Medium	25,26,27,28,29,30,31 32,33,34
Hard	35,36,37,38,39,40,41
Sub Total	24

General Comprehension

Stimulus - Target Language
Options - English

Difficulty Index	Item Number
Easy	42,43,44,45,46,47,48 49
Medium	50,51,52,53,54,55,56 57,58,59,60,61,62
Hard	63,64,65,66,67,68,69 70
Sub Total	29
TOTAL	70

BASIC COURSE FINAL EXAMINATION
Reading Comprehension Matrix

Translation

Stimulus - Target Language
Options - English

Difficulty Index	Item Number
Easy	1,2,3,4
Medium	5,6,7,8,9,10,11,12
Hard	13,14,15,16
Sub Total	16

Translation

Stimulus - English
Options - Target Language

Difficulty Index	Item Number
Easy	17,18,19,20
Medium	21,22,23,24,25,26,27,28
Hard	29,30,31,32
Sub Total	16

Contextual Comprehension

Stimulus - Target Language
Options - Target Language

Difficulty Index	Item Number
Easy	33,34,35,36,37
Medium	38,39,40,41,42,43,44,45,46
Hard	47,48,59,50,51
Sub Total	19

General Comprehension

Stimulus - Target Language
Options - Target Language

Difficulty Index	Item Number
Easy	52,53,54,55,56
Medium	57,58,59,60,61,62,63,64,65
Hard	66,67,68,69,70
Sub Total	19

TOTAL	70

The Reading Comprehension matrix for the Final Examination is also very similar to the Qualifying Test. We have eliminated the Vocabulary Identification section. Test length has been extended to seventy items, which will be reduced to sixty items following analysis.

The Final Examination for Speaking is a replica of the Qualifying Test counterpart except that test length has been approximately doubled. (See next page).

The provisions of the Achievement Test Program are being applied now to five basic courses under development in Arabic, Chinese, German, Korean, and Russian with Czech and Polish scheduled to begin this autumn. We plan our development program in five year increments and review that five year plan annually. For the next four fiscal years we will initiate basic course development for two languages each year, on average.

With twenty-eight language departments at DLI, it is evident that a number of years might pass before formal, full scale course and test development is begun for some languages which have a small annual student input. We do have some departments with the resources and initiative necessary to update their courses or tests. For these departments our Test Division provides orientation, training and technical consultation in upgrading their testing programs.

BASIC COURSE FINAL EXAMINATION
Speaking Matrix

Vocabulary Identification

Stimulus - Printed picture or cartoon with voiced TL question.

Response - Voiced TL description of picture item or situation.

Number of Items - 68

Structure

Stimulus - Complete printed English sentence. Incomplete printed translation of English sentence.

Response - Voiced complete TL sentence.

Number of Items - 60

Composite Communication

Stimulus - Situation printed and voiced in English.

Response - Voiced TL role based on situation.

Number of Situations - 7

Directed Response

Stimulus - Short story printed in English. TL questions printed and voiced.

Response - Voiced TL response to questions.

Number of Stories - 4

Number of Questions - 20

ED217732

COMPUTER-BASED ANALYSIS OF OPEN-ENDED FOREIGN-LANGUAGE TEST ITEMS

Torsten Lindblad (University of Gothenburg)

1 Introduction

In this paper I am going to describe a method of dealing with student-produced answers to open-ended foreign language test items. This method has been developed within the GUME/Tests project group in connection with research and test construction work carried out for (and with grants from) the National Board of Education and the National Office of Universities and Colleges. The problem and the solution to it will be seen from the consumer's point of view, i.e. the language teacher's and linguist's, not the computer specialist's. The theoretical and practical implications of the computer technique will be stressed, rather than the technique itself.

2 Background

In the field of language teaching there has been, in the last few years, a trend away from mechanistic drills with a stress on linguistic correctness, towards communicative exercises and creative use of language. This is true at least as far as Sweden is concerned and can easily be seen in textbooks and exercise books. This is, at least partly, influenced by trends in linguistic research with its growing interest in, among other things, psycholinguistics and sociolinguistics. The one prescribed norm, "Queen's English", has been replaced by concepts like acceptability and intelligibility, and this results in many different "Englishes". In educational evaluation there has, more or less simultaneously, been a shift of interest away from quantitative data towards qualitative information of different kinds.

I think the above facts - very briefly and roughly stated for reasons of space - explain what has been happening in the field of language testing over the last few years. Many of us have, no doubt, fallen out of love with the multiple-choice technique, at least when used indiscriminately and exclusively. It still deserves - and needs - to be used, but strict rules have to be adhered to and other formats must complement it.

Spolsky's three trends in language testing are too well-known to need more than a mention in passing. After the pre-scientific or traditional trend, which characterized much of the FL testing carried out in Sweden in the 1940's and 1950's, the psychometric-structuralist or modern trend followed with its heavy emphasis on reliability. The standardized tests used in Sweden and first developed in the late 1960's were created in this spirit and still largely reflect it. The third trend, the psycholinguistic-sociolinguistic or post-modern, is slowly gaining ground. It stresses validity and relevance and demands new test formats (See Appendix, Example 1).

FL013022

The questions that arise out of what has been said are these: what new testing techniques can we use and what methods are there of dealing with student answers in order to guarantee reliability while concentrating on validity? The computer techniques to be described in this paper represent one way of answering these questions.

3 Problems

The problems were first encountered in two large-scale research projects in the mid-1970's. In one of these, we wanted, among other things, to evaluate the FL proficiency of teacher trainees ready to start their teaching careers as compared to that of their future pupils. Difficult tests, largely consisting of open-ended items, were used. We soon found that a right/wrong scoring technique tended to hide as much information as it revealed. Not all answers which were outside the pre-prepared key were equally wrong. It was also very difficult to compare types of mistakes made by low-grade and high-grade students. Even if they both made mistakes, the question was: do they make the same kind of mistake? Scoring hundreds of items from hundreds of students by hand and producing lists of dozens of variants to one and the same item was not manageable, and we realized that a lot of information was lost.

We had the same experience in an even larger project evaluating the proficiency of pupils from grade 5 to grade 12, i.e. eight successive age levels, in some cases parallel streams etc. Complicated scoring keys (which will be described briefly below) were produced. Since many scorers had to be used over a long period of time for more than 10,000 answer sheets, each with 250 items, all of them of the open-ended format, both inter-rater and intra-rater reliability turned out to be far from satisfactory. The scoring keys had to be produced as the scoring work proceeded on a day-to-day basis. The inter-subjectivity that was, at best, achieved could not be questioned afterwards and changes introduced. Could not a technique be found whereby lists could be produced of all student responses, whether right or wrong, so that a group of experts could discuss scoring principles and the scoring itself be carried out afterwards? Could separate lists be made for different groups of students without unreasonable extra work? Could the effects of different scoring principles be compared and correlated? Could the lists produced for successive age cohorts be used to study changes of the interlanguage of the students even if full correctness had not been arrived at, i.e. changes that would be completely hidden if right/wrong scoring were used? Those were the questions we asked ourselves.

The project was also involved in the production of standardized tests in English as a foreign language (and later in French and German) for different age levels. Three new tests in English are constructed every year and have to be carefully pretested. There were routines to deal with the multiple-choice items, but at the National Board of Education a number of people had to spend the whole summer listing pupil answers. In these lists the various answers appeared in the random order in which they were found. The number of pupils who made each error was calculated, but there were, quite naturally, mistakes, and no point-biserial correlations could be calculated. This yielded a less satisfactory basis for judging open-ended items as compared with multiple-choice items. The types of items used were short passages or mini-dialogues in which there were gaps where one word or several words (usually 2-4, never more than 6) had been deleted. These will be referred to as one-word gaps and word-string gaps respectively. Lately, the cloze technique with variable-ratio deletion has been introduced (referred to as "cluze test").

Work on these new testing techniques soon turned out to be difficult, not to say impossible, without some kind of computer-based method of item analysis. We needed to know what different answers had been given (in order to decide whether to use the item or not and to produce keys for the teachers) and what the discrimination indices were. We also needed to be able to study how an item functioned among students of different ability and proficiency.

4 Traditional item analysis

The item analysis techniques used to study multiple-choice items yield what is often referred to as facility and discrimination indices. This means that the percentage of the student group who answered correctly is calculated together with the point-biserial correlation, i.e. the correlation between the item and the test total (preferably not including the item itself). In the item analysis program that we have constructed within the project, the percentages of the pupils who have chosen each distractor are also calculated as well as the point-biserial correlations for the distractors (Appendix, Example 2). These data tell us whether a multiple-choice item has worked properly. Exactly the same program can be used for open-ended items which are first scored according to a pre-prepared key. The correct answers can then be considered as alternative A and all incorrect answers as B. Descriptive statistics can also be calculated, which gives us means, standard deviations, reliability coefficients etc. for the test total or any subtotal that we want.

Multiple-choice answers can be fed into the computer quite easily with a variety of techniques which will not be discussed here.

5 Preliminary solutions to the problems

The solution to the problems and questions mentioned above was the invention of an A-B-C-D scoring scale. Different answers were not weighted with more points for a completely correct answer and less points for different degrees of incorrectness. The scale was a nominal scale: correct answers were marked A, incorrectly spelled variants of A (and C) words were marked B, correct or acceptable variants or synonyms C and incorrect answers D. An E was used when no answer at all had been given (Appendix, Example 3). These different letters were then fed into the computer item by item for all students. The ordinary item analysis program could then be run, yielding percentages and point-biserial correlations for each category. It was also possible to add the A's and B's or the A's, B's and C's, and thus arrive at a dichotomous scoring: correct or acceptable answers, whether correctly or incorrectly spelled, as opposed to completely unacceptable answers. Each student could be given two or three different sums according to what scoring principle was used and these sums could then be correlated. It may be mentioned, in passing, that these correlations tended to be around .98 or .99, which means that the addition of "acceptable" variants makes very little difference (at least for the advanced students that were studied), except that all mean scores are raised a little. There were, however, greater differences among low-proficiency students than among those of high proficiency.

There are problems connected with this technique. The scoring has to be done by hand by fairly well qualified persons, keys consisting of long lists of variants have to be produced and adhered to, and the marking, once done, cannot be changed. Corrections of mistakes discovered late or a change in overall marking policy cannot be introduced.

6 "PRINS"

The solution to all these problems was PRINS, the "provinmatningssystem", i.e. the system of computer programs for the 'feeding-in' of tests and their statistical processing. This system of computer programs was constructed by the GUME/Tests project group computer specialist and programmer Jan-Gunnar Tingsell in cooperation with me. I represented the consumer point of view. It has now been used for three or four years and has been gradually revised into its present form.

The feeding-in procedure. A secretary, who needs only a couple of hours of "computer conversation" training, sits at a video terminal, i.e. a fairly ordinary type-writer keyboard with a video screen. The program is interactive, which means that the computer asks questions that must be answered and reacts when certain types of errors are made. Corrections of typing errors are easily dealt with. This means that few unwanted mistakes are introduced at this stage. On the other hand, the typist must, of course, copy all the mistakes made by the students. (The typists have to have other things to type also so as not to get completely confused as far as spelling is concerned!) When several words belong to the same item a full stop is inserted between the words of the item, e.g. "he.came.home". In French, the accent marks etc. have to be replaced by figures, which makes the spelling look somewhat strange, but with some practice the words can quite easily be deciphered: l'été becomes "l'e2te2", où is typed "ou3", the circumflex is 4, ça becomes "c5a" and Noël "Noe6l". After the computer has been told how many items to expect from each student, it will react if too few or too many answers are typed. If the pupil has given no answer, an 0 must be inserted. The machine will also react if the identification number does not conform to the standards first given.

The student answer sheets are thus dealt with quite mechanically by a typist who does not need to know what the expected answers are. In this fashion, all pupil answers are easily and rapidly fed into the computer for further analysis.

Listing and scoring. After all pupil answers have been typed into the computer, the typist can ask the program, by calling a different subroutine, to produce a printed list of all student answers item by item. The machine then arranges the answers in alphabetical order (Appendix, Example 4) with capitalized words at the end; German nouns (as well as English words like German or Tuesday) will thus be separated into two groups: Those which have been spelled correctly with a capital letter and those which have not. In the list, the numbers of occurrences of each answer as well as these numbers in per cent of the student group are also given. The answers in the list are given running numbers which can be used to refer to them.

The next step in the procedure is the feeding-in of the scoring key. For each item, one or several answers in the list are marked as correct by entering their numbers as given in the list. A new list can now be produced in which the correct answer or answers are marked with an R. This then forms the basis for an ordinary item analysis in which all R-marked answers are treated as alternative A and all other answers as alternative B. In addition, the number of blanks are given. In this step, which is taken by the computer automatically and which can therefore be handled by the typist without the help of a computer specialist, facility and discrimination indices are given together with the ordinary descriptive statistics data mentioned above for multiple-choice items. The answers produced by each pupil are now compared by the computer with the key and the pupil awarded a test total, which we can refer to as "sum A", i.e. the total number of correct answers according to a strict, pre-prepared key. The answers to the items have now been dichotomized and are treated as if the items were ordinary two-choice tasks. The lists can be used in the different ways that will be referred to below.

Scoring principles. The test constructors can now discuss the various answers and decide on a new key. Incorrect spellings may be overlooked, different criteria can be applied (e.g. if the spelling produces another existing English word the answer is not accepted, if the spelling implies uncertainty about the pronunciation of the word it is not accepted etc.). A large number of synonyms, near synonyms or strange spellings may be entered as "correct" answers (Appendix, Example 5). Then the whole procedure is run again, producing new and, of course, higher test totals for the pupils, "sum B", and a new item analysis is performed. In this second run of the program a large number of answers may very well be treated as alternative A. It may be added that it is quite common to have lists of more than fifty variants for one-word gaps and over one hundred for word-string gaps. The maximum to date, I think, is 218 different answers to a four- or five-word gap. In cases like that more than twenty answers may very well have to be accepted.

The two test totals, sums A and B, can be entered as pupil variables and correlated. The correlation coefficients are very high even though they vary for different groups of students.

7 Areas of application

The PRINS computer program system is complex and has taken a number of years to develop into its present form. It is related to the work carried out within the field known as computational linguistics and is, I would think, of some interest to computer specialists. For the language teacher, linguist, and test constructor other aspects of the system will be worth mentioning. I will discuss some of the ways in which we use and plan to use PRINS.

Test construction. As was mentioned above (Background) the open-ended item testing technique is being used more often than was the case a few years ago. This is in line with a trend to stress the importance of student production. The short-answer gap format lends itself to item analysis techniques similar to those used earlier for multiple-choice items. The PRINS computer programs yield lists of all student responses produced. With the help of this system, it is possible to produce these lists before any scoring is done, and afterwards to score in several different ways and study the effects of these (Appendix, Example 6). If new scoring principles are decided on, new results for all students can be calculated in a few minutes by a secretary with little or no computer training. The item analysis results are then used in the ordinary way for the selection of items; items which are considered too easy or too difficult or which discriminate too poorly are rejected, those with suitable facility and discrimination indices are accepted. Scoring keys for the teachers who are going to use the tests can be produced on the basis of the lists of responses.

After the test constructor has decided on which items to include in the test, new descriptive statistics can be produced for this set of items, which will give a fairly good idea of how the final version of the test will function. This prognosis can, of course, quite easily be tested against the sample of tests collected to produce standardized norms for grading.

Discussions of "acceptability". The lists of answers can be used in discussions of scoring principles. Which answers can be accepted, where do we "draw the line"? Some items also turn out to be too "open", i.e. too many quite different answers have to be accepted for semantic reasons. In some cases a fairly great variety can be accepted (e.g. in cloze tests where reading comprehension is of central interest), whereas in other kinds of items (e.g. if one wants to test specific words or grammatical structures) it is less acceptable if the student is able to "escape" by using a very easy word or phrase. I do not intend to discuss these various principles and points of view here, I only wish to stress the fact that the PRINS technique is very useful in producing a background for this kind of discussion.

Cloze testing. I would like, however, to point out here that we have found in Sweden, quite recently, that the cloze technique gives rise to a large number of scoring problems which, for some reason, never seem to have been discussed in the language testing literature. In Oller's *Language tests at school* (1979), where the cloze test is described and discussed at some length, only the exact-word and the acceptable-word principles are mentioned. How thirty different spellings of "necessary" (see Example 3) or a similar number of variants of "languages" should be dealt with, is not touched upon. I would like, in some other context, to come back to this problem which is, at this very moment, being discussed in Sweden very intensively. The question arises: what are we actually testing? There is an obvious risk that the cloze test becomes not an integrative test with a stress on reading comprehension, but simply a spelling test, and a rather uneconomic one at that.

Research. The PRINS technique can also be used for other purposes in language teaching research. If hundreds of answers from hundreds or thousands of students are to be analysed, PRINS is an excellent aid. It will yield lists of correct and incorrect answers for any subgroup of the student population under study (Appendix, Example 7). Different mistakes, which reflect different stages in student interlanguage, can be studied for various groups of pupils, e.g. pupils at different age levels. We can make, on the basis of the lists, qualitative analyses of the proficiency of our pupils. In an analysis of student-produced varieties of the past tense question "did he buy" (see Example 4), we can find out whether the students use the 'do-construction' at all, whether they use any past tense markers etc. - Some of our large-scale evaluation studies were carried out before the PRINS program existed, but we continually use the system for minor studies of this kind. It is also a valuable technique which can be used to analyse diagnostic tests.

8 Availability

The program can be put at the disposal of anyone interested at nominal cost (exact figures in pounds and pence can be given on request), as long as the computer system is compatible with our IBM routines. Questions can be answered by me or the project programmer, Jan-Gunnar Tingsell.

It should be mentioned that the PRINS program interacts in Swedish, but a version with all questions and instructions in English will be available. There is also a user's manual which explains in great detail how to use the system.

9 Summary

I have tried, in this paper, to describe and discuss some ways in which the PRINS computer program can be utilized in test construction and linguistic analysis. We do not understand now how we could ever do without these programs. We hope friends and colleagues abroad will be able to benefit from our work and are looking forward to an exchange of ideas and findings.

Appendix

Examples of test items, item analysis data, lists of student responses etc.

Example 1: Examples of items that demand student production

"One-word gaps"

1. I should like to go back to the town _____ I spent my childhood.

2. "What? Back already? Didn't you _____ yourselves?"
   "Oh yes, we had a wonderful time, but the party broke up early."

3. Uncle Joseph keeps talking about himself. I'm tired of _____ to his endless stories about how clever he is.

"Wordstring gaps" (2-4 words, max 6)

4. The situation has now become intolerable and _____ done about it immediately if we are to avoid a major catastrophe.

   (Acceptable answers: something has to/must/ought to/should be
   we have to/ must/ ought to get/have something)

5. PAT: Mr. Nevinson seems to be quite well off.
   TOM: ____________ for a living?
   PAT: He's an engineer.

6. "No one believes in ghosts nowadays, ____________?".

"Cluze test" (cloze with variable-ratio deletion and pre-tested items)

Nigel stood in the bedroom, holding his breath. The large country house was full of odd creaks during the night, but the sounds he had just heard were different. Carefully, he opened the door of the wardrobe, chose a dark blue dressing-gown and put it on. Very quietly, he slipped out of the room and stared into the darkness. Silence.

Just as Nigel was beginning __7__ think that he had only imagined the __8__ , he heard them again, faint and metallic. He tied the cord __9__ his dressing-gown, and __10__ downstairs. He tiptoed across the large hall. In the drawing-room he stopped again __11__ listen. Yes, the noises __12__ from the library. He looked in....

Example 2: Examples of multiple-choice items with item analysis data

1 I'm underpaid at the moment, but they've promised me a rise in ______ from next January.

A. money
B. salary
C. economy
D. loan
E. paying

		easy course			advanced course		
		%	r_{pbis}	score +/	%	r_{pbis}	score
A		14.8	-.14	8.8	8.3	-.27	12.1
B	R	52.4	.49	11.8	64.9	.47	17.6
C		4.2	-.15	7.4	4.4	-.19	12.0
D		3.7	-.13	7.4	2.0	-.19	10.0
E		24.3	-.35	7.7	20.5	-.21	14.2

+/ out of 25 items

2 The man seemed to be dead when they pulled him out of the water, but they gave him artificial ______ and after a while he opened his eyes.

A. breathing
B. first-aid
C. respiration
D. ventilation
E. support

		%	r_{pbis}	score	%	r_{pbis}	score
A		52.9	.44	11.6	76.6	.24	16.7
B		36.5	-.36	8.2	12.7	-.18	14.0
C	R	6.9	-.13	8.2	9.3	-.10	14.7
D		3.7	-.07	8.7	0.5	-.11	9.0
E		0.0	-	-	1.0	-.08	12.5

/of the above two examples, item 1 was accepted and item 2, of course, rejected - the terms 'easy course' and 'advanced course' refer to the two alternative EFL courses the students can choose between in grades 7 - 11 of the Swedish school system/

Example 3: Examples of "categorized" scoring of open-ended items

- Do I have to do it tomorrow?
- If you can, but it's not absolutely __1__ .

A = correct word according to key
B = correct word, wrong spelling
C = correct synonyms (no examples in these items)
D = incorrect words

A necessary
B nececcary nessesary nessecary nessecery neccesary
neccessary necessery nessary necceserry nessisary
nessesery necesary nesecarry nescassary nesessary
necessairy nicessery nescesary nicissary necesarry
necesery nesseccerry nesserceary nessessairy necassary
neassasary nessessary nessicery neccessery nesesarry
nesesary nessercery nesserrary
C -
D neccecerely nesserserly urgent
necessarely nesseserly
nesserarely sure

- It's very cold today. Have you checked the temperature?
- Yes, it's 20 __2__ below zero.

A degrees
B degreades degres degreeds degreas degreze degress
degrays degrece degrades degreeze deegres degrease
degreez degreeth degreece degreese degrêe digrees
degreads degreace degrese
C -
D celsius degre grads marks
centdegree degree graduates points
centegrads grades grees sight
centigrades grade's grez stresses

Example 4: Typed version of the computer list of the first 39 suggested 'answers' to item in which "did he buy" was the expected response

ITEM 17	R/F	NO	ANSWER	FREQU	-%-
		1	ABOUGHT	2	0.34
		2	ARE BOUGHT AND	1	0.17
		3	BAYS HE	1	0.17
		4	BAY	1	0.17
		5	BOUGHTED	1	0.17
		6	BOUGHTS HE	3	0.51
		7	BOUGHT HE ON	1	0.17
		8	BOUGHT HE	2	0.34
		9	BOUGHT	3	0.51
		10	BOURGH	1	0.17
		11	BOURGTH	1	0.17
		12	BUY	4	0.68
		13	BY	1	0.17
		14	DES HE BOUGHT	1	0.17
		15	DID HE BAUGHT	1	0.17
		16	DID HE BAY	17	2.90
		17	DID HE BOUGHT	129	22.01
		18	DID HE BRING	4	0.68
		19	DID HE BROUGHT	8	1.37
		20	DID HE BUING	2	0.34
		21	DID HE BUYING	2	0.34
	R	22	DID HE BUY	168	28.67
	R	23	DID HE BUY?	1	0.17
		24	DID HE BYE	7	1.19
		25	DID HE BY	12	2.05
		26	DID HE WRITE	1	0.17
		27	DID HI BOUGHT	2	0.34
		28	DID YOU BOUGHT	1	0.17
		29	DOES HE BAY	3	0.51
		30	DOES HE BOUGHT	28	4.78
		31	DOES HE BOY	1	0.17
		32	DOES HE BROUGHT	1	0.17
		33	DOES HE BUY	16	2.73
		34	DOES HE BYE	3	0.51
		35	DOES HE BY	1	0.17
		36	DOES HE BOUGHT	1	0.17
		37	DOES	1	0.17
		38	DOSE HE BOUGHT	1	0.17
		39	DO HE BAY	1	0.17

Twelve of the "answers" to this item can be arranged in the following order for "qualitative evaluation" of student interlanguage, thus:

bay he
buy he
bays he
bought he
boughts he
boughted he
do he bay
do he bought
did he bought
does he bought
does he buy
did he buy

Example 5: List of answers with correct and acceptable answers marked R and A respectively (see Example 6, item 10 below)

			advanced course	easy course
	1	at		.4
	2	be		.4
	3	bird		.4
	4	dressing-room		.4
	5	from		.4
	6	get	3.2	9.5
	7	gett		.4
	8	goe		.4
	9	goes	1.2	4.8
	10	goe's	.4	
	11	going	2.0	2.2
	12	gone	.4	.4
	13	go	2.8	16.0
	14	go's		.4
A	15	got	1.2	.4
	16	last		.4
	17	look	.4	.9
A	18	looked	.4	.4
A	19	ran	.4	.4
	20	run		3.5
	21	set	.4	
	22	sett		.4
A	23	Slipped		.4
	24	sound		.4
	25	standing		.4
	26	step	.5	.9
	27	steped	.4	
	28	stepped	1.2	
	29	stept	.4	
	30	there		.4
A	31	tiptoed	.4	
	32	walk	3.6	5.2
	33	Walk	.4	.4
R	34	walked	7.1	3.5
A	35	Walked	.4	
	36	want		.4
A	37	wen		.4
A	38	wendt		.4
R	39	went	69.8	36.4
A	40	Went	.4	.4
A	41	wennt	.4	
	42	which		.4
A	43	wolkd		.4
	44	wont		.4
	45	0	2.0	6.1

Example 6: Examples of the effects of different marking principles (from "cluze" test; see Example 1).

... he had only imagined the 8 , he heard them again.

Correct answers according to pre-prepared key = noises, sounds
Accepted variants = noices, noicees, saunds
Incorrect answers = noise, sound

Easy course		Advanced course		
%	r_{pbis}	%	r_{pbis}	
21.6	.40	48.8	.50	(correct)
23.4	.38	69.4	.49	(correct + accepted)

He tied... and 10 downstairs.

Correct answers according to pre-prepared key = walked, went
Accepted variants = got, looked, ran, slipped, Went, wendt, wen, wolkd, wennt

Easy course		Advanced course		
%	r_{pbis}	%	r_{pbis}	
39.9	.56	77.0	.47	(correct)
43.3	.55	80.6	.56	(correct + accepted)

Example 7: Examples of answers produced by students in the advanced course ("special course") and the easier course ("general course") in grade 11.

Dennis is two years older ______ me.

Advanced course		Easy course	Easy course
than	88%	than (70%)	for
then	12%	then (17%)	of
		and	on
		an	that
		at	when

Pages 88-94, which contained an article in German, have been omitted.

ED217733

TESTING ORAL PROFICIENCY IN LIFE-LIKE SETTINGS: RATING SCALES AND RELIABILITY

Jan Mets (Cito, Arnhem)

In 1978 the Dutch National Institute for Educational Measurement (Cito) started a project aimed at the development of oral proficiency tests for German, French and English, the three foreign languages most popular with pupils in secondary education.
The early struggles were reported on by Jan van Weeren at the IUS meeting in Hürth (1979) and published in Klein-Braley, C. and D.K. Stevenson, (1981), (eds), Practice and problems in language testing 1. (Verlag Peter D. Lang). I intend to continue where he left off. Before I do so, however, I think a short reminder of the main characteristics of the tests developed so far and of the first experiment carried out in early 1979 and its results might come in useful.
For the three languages mentioned, tests were constructed to find out how well, after about three years of foreign language instruction (at a rate of three periods a week) pupils cope, linguistically speaking of course, in everyday real-life settings. Each test requires the pupil to play the part of someone entering into contact with a native speaker of the language concerned. This latter part is played, as one might guess, by his teacher, who occasionally, however, is supposed to combine his acting with the role of examiner, i.e. when he has to set his pupil a specific assignment. Let me give you an example.

During a cycling-tour in Britain you and a friend stay at youth-hostels. One day you discover your wallet is missing. You go to the police station.

1 *Good morning, what can I do for you?*
Explain your problem.

2 *Could you tell me exactly what was in your wallet?*
Describe the things in your wallet. (picture 1)

3 *Could you tell me when you first missed your wallet?*
Say what you were doing. (picture 2)

(picture 1 shows some money, a boat ticket to Holland and a passport; picture 2 shows someone running along the platform of a railway-station)

The two different types of print indicate to the teacher which role he is supposed to assume: the examiner's lines are in roman, the italics serve as a cue for him to try and speak in a policeman's voice.

FL013023

From this example you can also see what we understand by everyday real-life settings, as from the titles of ten tests published so far:

1 At the railway station
2 In a train compartment
3 Cycling
4 Hitch-hiking
5 Camping out
6 Holland
7 Shopping
8 Seeing a doctor
9 At the police station
10 Public Holidays/Festivities

Strictly speaking, i.e. if we make the distinction Dr van Ek made in his The threshold level between settings and topics, two of these titles, 'Holland' and 'Public Holidays/Festivities', indicate topics rather than settings. Each test consists of ten assignments. Immediately after each of these, the pupil's performance is assessed according to the categories on the rating scale. Originally this rating scale featured these categories:

A1	No answer forthcoming/answer completely meaningless in context	.5
A2	Speech is so halting as to interfere with intelligibility	1
A3.1	Minor mistakes are made, not interfering with intelligibility however	1
A3.2	One or two basic mistakes are made that do not affect intelligibility	2
A3.3	One or two mistakes are made that do affect intelligibility	3
A4	Gross errors of pronunciation are made	1

The figures on the right represent the number of points deducted in each case. I should add here perhaps that the first category ('No answer forthcoming' etc.) excludes the other categories and that the sub-categories of A3 are also mutually exclusive.
Ten teachers of German were requested to assess the taped responses of 60 pupils. Each test consisted of 17 items. Here are some of the findings.

Agreement among raters and reliability of the German test

$R_i = .47$ Overall agreement among raters (Intraclass R)

A1 $R_i = .41$
A2 $R_i = .14$
A3 $R_i = .43$
A4 $R_i = .16$

Agreement among raters for each category

A3 K_0 = .31

A3.1 K_1 = .09

A3.2 K_2 = .20 Agreement among raters for each of the sub-categories

A3.3 K_3 = .19 of A3

r_2 = .73 Reliability of the test as such

These figures and especially the low ones for categories A2 (halting speech), A4 (errors of pronunciation) and sub-category A3.1 (minor mistakes) led to a regrouping and redefining of categories and consequently a new rating scale:

LEERLINGNUMMER

OPDRACHTEN

Du fährst allein per Anhalter in den Süden. (Bild 1) Abends meldest du dich auf einem überfüllten Campingplatz an. Du bist müde und möchtest recht gern schlafen gehen.

SCORING	antwoord goed begrijpelijk; incidentele fout(en)	antwoord moeilijk te begrijpen	antwoord blijft uit/ volkomen onbegrijpelijk in gegeven context	
1.	☐	☐	☐	1. Beim Empfang triffst du eine Dame an. Sie sagt: "Guten Abend! Sie wünschen?" Was antwortest du?
2.	☐	☐	☐	2. Die Dame sagt: "Es ist kein Platz mehr frei." Du läßt dich aber nicht abschütteln. Versuche, sie zu überreden, indem du ihr das Format deines Zeltes beschreibst! (Bild 2)
				..........etc.
16.	☐	☐	☐	16. Er sagt: "Es gibt ein Freibad in einer Entfernung von etwa fünf Kilometer. Ich fahre in zehn Minuten dorthin!" Das paßt dir gut! Was sagst du?
17.	☐	☐	☐	17. Der Nachbar sagt: "Aber natürlich!" Du willst dir zuvor im Laden noch etwas besorgen und mußt dich deshalb beeilen. Verabschiede dich für kurze Zeit!

spreekgemak	goed	
	voldoende	
	onvoldoende	
	slecht	

uitspraak	goed	
	voldoende	
	onvoldoende	
	slecht	

There are only three categories on paper, though actually there is a fourth: if a pupil's achievement is perfect or nearly perfect this need not be recorded as no points are deducted. The definitions of the three remaining categories read as follows:

B_1	response is quite intelligible and appropriate, even if some minor mistakes are made	1
B_2	response is intelligible and appropriate but takes some effort to understand	3
B_3	response is not forthcoming/is not intelligible/ not appropriate in the context	5

It is emphasized in the instructions for raters that when they have to ask themselves whether a response is intelligible they should do so from the point of view of a native speaker of the language concerned. The categories are, of course, mutually exclusive.

At the bottom of the rating-scale and so at the end of the test, the rater can record his evaluation of the examinee's fluency and pronunciation by assigning them to any one of the four categories: good, adequate, inadequate, weak.

Thirteen teachers of German were invited to try out this new rating-scale in assessing the same taped performances of 50 pupils as were used on the earlier occasion. The figures for agreement and reliability this time were:

$R_1 = .69$

$K_u = .43$

$K_1 = .28$

$K_2 = .21$

$K_3 = .33$

$\alpha_2 = .77$

Agreement among raters was measured by means of intraclass R : $R_1 = .69$, a considerable improvement on R_1 for the earlier model as this was .47.

Fleiss' kappa was again used to measure agreement among raters in assigning responses to each of the categories and from these results it can be concluded that raters disagreed most about assignment to category B_2 (response takes some effort to understand).

As far as fluency and pronunciation were concerned, the results remained below par. Raters appear to have most difficulty in distinguishing between good and adequate achievements on the one hand and inadequate/weak achievements on the other.

As in the case of the earlier model, the reliability α_2 was measured. This was .77, quite a satisfactory value for only 17 items.

Another improvement of the later rating-scale concerns differences between achievements of raters during the first half of the rating session and the second half.
On the earlier occasion it appeared that, whereas raters themselves showed greater consistency in their assessments, agreement among raters grew less. With the later model both the consistency of raters themselves and agreement among them increased. From this fact it can also be concluded that it might be useful to try and find out what effect training could have on raters. Finally, if the phenomenon of systematic differences among raters is not regarded as an undesirable quality and so the factor of bias is eliminated, it appears that with the first model agreement among raters increases much more than with the second model:

first model: $R_1 = .47$ $R_2 = .64$ → an increase of 36%

second model: $R_1 = .69$ $R_2 = .78$ → an increase of 13%

This means, of course, that bias plays a less important role in the case of the second model.

When these results had become known it was decided that the experiment would be repeated for French and English and that the rating-scale (having produced satisfactory results on the whole) would only be adapted in one respect: the categories for fluency and pronunciation would be reduced in number from four to three: good, adequate and inadequate.
The same procedure was used as for German. Ten to twelve experienced teachers of English and French were invited to assess 50 taped achievements of pupils on two tests i.e. twenty responses per pupil. There was only a brief introduction, the main point of which was to make the teachers familiar with the rating-scale as this was something new to all and assessments had to be made in the relatively short gaps of four seconds in between responses. We deliberately avoided any discussion about standards of evaluation at the beginning of the rating-session as we were interested in finding out the merits (or demerits) of the rating-scale when used by the average teacher. For the same reason we did not stress our own point of view in the matter at all, but confined ourselves to presenting the achievements of two pupils on ten items each, together with our own assessments and a very brief explanatory comment.

The agreement among raters was once again measured by means of intraclass R - on the basis of total scores - and Fleiss' kappa for each category. The values found were, in the case of French raters, quite similar to those found in the earlier experiment, but the English figures lagged far behind:

Agreement among German, English and French raters

	German 1979	English 1981	French 1981
Intraclass R total scores	.69	.42	.68
Fleiss' kappa	K_u = .43	K_u = .31	K_u = .51
per category	K_1 = .28	K_1 = .18	K_1 = .35
median value	K_2 = .21	K_2 = .14	K_2 = .18
	K_3 = .33	K_3 = .26	K_3 = .42

The reasons for the less satisfactory results for English are presumably the fact that the selection of pupils' responses presented to the raters showed too little variety as far as the quality of achievement is concerned and the fact, related to the former, that assignments were fairly easy. From the kappa values for the categories in the cases of French and German it can be concluded that the second category ('response is intelligible and appropriate but takes an effort to understand') posed most problems to the raters.
Reliability (α_2) was satisfactory for French, as it was for German, but rather poor in the case of English. (Further analysis of data confirmed our assumption that the assignments were rather easy for the pupils concerned) As to the categories 'Fluency' and 'Pronunciation' once again raters proved not to agree with each other to any acceptable degree judging from the K-values for each category as a whole (K_t) and for each sub-category. I present here those for French:

fluency	K_t = .27	pronunciation	K_t = .19
	K_g = .39		K_g = .28 (good)
	K_v = .12		K_v = .11 (adequate)
	K_u = .34		K_u = .17 (inadequate)

These results are not very satisfactory. And so, considering that they constitute assessments made only once, after the actual test is over, it was decided that not too much weight should be assigned to them, not so much anyhow as had been done up till then. In our latest design an examinee's fluency and pronunciation will cause him to lose a few points if they are exceedingly bad and some bonus points will be awarded if they are exceptionally good. Whether this will work any better can be seen perhaps when the results of the next rating session are analysed.
I have already mentioned the possibility that training of raters might have a positive effect. As agreement among the English raters had been so much lower than what their colleagues for French and German had demonstrated, it was thought that here in particular careful instruction might show positive results. If so, it could be decided that the same sort of instructions would be provided to all future users of our

oral tests. For this reason the instructions that we were going to give during another rating-session would have to be in a form that would be practicable later. To arrive at as profitable instructions as possible an attempt was made to try and discover the causes of disagreement among the English raters. First the agreement among raters on each item was measured by means of intraclass R as well as the item difficulty, which here takes the form of the mean negative score of pupils:

ENGLISH

Intraclass R for each assignment		diff. of each assignment
1	.39	1.11
2	.28	.90
3	.46	.95
4	.23	1.39
5	.24	.78
6	.18	.75
7	.18	.73
8	.31	1.38
9	.25	.94
10	.15	1.03
11	.36	.36
12	.34	.20
13	.37	.52
14	.37	1.51
15	.20	.69
16	.27	.51
17	.21	1.57
18	.22	.69
19	.33	.51
20	.40	.99

Secondly, those assignments were selected which combined a low value for agreement with a very low or very high figure for difficulty. Hypotheses concerning grounds for disagreement were formulated, after which the pupils' responses were listened to once again to see if our hypotheses could stand. Let me give an example. One of the assignments we had a closer look at was number 17 ($R_1 = .21$; diff. = 1.57), which reads: *We have many foreigners living in Britain. What's the situation like in Holland?* Though it should be possible to give a satisfactory answer in one sentence, we supposed that pupils might have problems here if they let themselves be tempted into lengthy observations, mixing accounts of colonial pasts and personal opinion. Their main problems would be forcing their reflections into idiomatically and grammatically correct forms. Raters, we further supposed, would be tempted by these grammatical and idiomatical defects to assign intelligible responses to a lower category than they actually deserved (for instance in the case of a pupil referring to a large group of foreign workers in Holland as 'Turkeys').

We found that almost 70% of the responses analysed were intelligible in spite of certain grammatical or idiomatical errors and so it can be concluded that a number of raters did attach greater importance to formal correctness than to intelligibility of the responses. We also found here that about 25% of the pupils gave very short answers, which, as one might guess, is a recurrent problem in assessing productive achievements and not al all specific to this assignment. Our general conclusion here was that if pupils take their time in finding the suitable words to express more complex phenomena or, these words failing them, work their way around them without sacrificing too much of their intelligibility, their achievements do not deserve to be assigned to categories 2 or 3.
This procedure, in which nine assignments were analysed, led to a number of conclusions and suggestions with a view to the training of raters, the construction of future tests and the definition of criteria for assessments. Here I shall only mention those with a bearing on the training of raters:

1 If a response displays a pupil's defective knowledge of notions it is not therefore categorically unintelligible. It could be assigned to cat. 2 ('intelligible/appropriate, but taking an effort to understand') or even cat. 1 ('quite intelligible even if some minor mistakes are made'). Consider for example: 'Could you please pass me the thing over there, to put my cigarette ashes in.'

2 If a response displays a lack of good manners (from a British point of view) it should be assigned to category 2 (see also suggestion (5) for an adapted definition below).

3 A response that is correct from a grammatical and idiomatic point of view but does not comply with the assignment should be assigned to category 3 ('response not appropriate') If for exemple the assignment is, as in no. 10 of the English test, 'Explain why you hesitate', hesitation should be expressed in one way or another.

4 The length of a response or the amount of information supplied are not in themselves sufficient grounds for assignment to any category. If a response, even though very short, is intelligible and appropriate it should be assigned to category 0 or 1.

5 The definition of category 2 should be extended to include responses that are less acceptable from a social point of view.

On the basis of these conclusions a programme of instructions for raters was compiled and taped. In October 1981 the same group of raters that assessed the 50 achievements on the earlier occasion will be asked to attend another rating session. This time they will receive careful instructions by way of the taped programme referred to. It includes two successive possibilities for them to compare assessments they have made themselves of samples of pupils' achievements with our proposals for evaluation plus accounts for these. Once again there will be no general discussion before the actual rating is done as we are interested in finding out what effect such taped instruction can achieve. The raters' achievements this time in assessing the responses of pupils 1-25, will be compared to their earlier achievements on the same material. It should be added that this can be done without any misgivings, as a comparison of the figures for agreement among raters on the first batch of responses and the second one showed there was no learning effect. At the same time the new design of the categories for 'fluency and 'pronunciation' will be put to the test, as I have already mentioned. In this way, it is hoped, a contribution will be made to finding a method of testing oral proficiency in a more reliable way.

ED217734

FL013024

THE TEST OF ABILITY TO SUBORDINATE: PREDICTIVE AND CONCURRENT VALIDITY FOR ATTAINED ESL COMPOSITION

Kyle Perkins (Southern Illinois University)

Currently there is wide-spread belief that there is a strong (almost causative) relationship between syntactic enhancement and overall writing proficiency. A further claim exists that improvement in syntactic elaboration will be concurrent with improvement in attained writing proficiency. In particular, a number of studies have suggested that oral or written sentence-combining enhances the syntactic maturity of elementary and secondary native-speaker school children (cf. Miller and Ney, 1968. Hunt and O'Donnell. 1970; Vitale, King, Shontz and Huntley. 1971; Obenchain. 1971; O'Hare. 1973; Fisher, 1973; Perron, 1974; Combs. 1975; Pedersen. 1976; all cited in Kerek, Daiker, and Morenberg. 1980). Syntactic maturity in these cases refers to elaboration and expansion within sentences as promoted by sentence-combining exercises. The prime concern of the classroom composition teacher, however, is not syntactic maturity per se, but whether gain in syntactic maturity correlates with gain in overall writing ability.

O'Hare's research did indicate that there was a correlation between improved syntactic maturity and attained writing proficiency, while Combs' and Pedersen's study further corroborated O'Hare's initial findings. Kerek, Daiker and Morenberg's study involved a tightly-controlled, large-scale study of the effectiveness of sentence-combining on freshman composition at Miami University of Ohio. They found that the experimental groups which received sentence-combining exercises showed significant gains in mean clause length, words per T-unit, holistic evaluation and forced-choice rating-measures of attained writing proficiency.

Although the effects of sentence-combining exercises are of great interest to the composition research community, they are not the focus of this paper. Instead, this paper reports the results of a study in which the following research questions were addressed: i) whether there is a strong relationship between syntactic enhancement, as measured by TAS, and overall writing ability; ii) whether a subordination test, supposedly measuring a student's ability to embed sentences, has any predictive validity for attained writing proficiency; iii) whether the subordination test mentioned in ii) has any concurrent validity; iv) whether different scoring methods affect the correlation coefficients?

1 Procedures

1.1 Subjects

The subjects for this study were adult foreign students taking

Freshman Composition for Foreign Students in the Department of Linguistics at Southern Illinois University; one section of 14 students was used for this study. Nationalities represented were Japanese, Malaysian, Chinese, Spanish, and West African. The subjects followed the regular composition syllabus, and they did not receive any sentence-combining exercises. The term was 15 weeks long.

1.2 Materials

At the beginning of the term, the subjects were given Davidson's (1978) Test of Ability to Subordinate, an instrument which tests students' abilities to subordinate in writing. TAS requires that students combine two or three sentences into one sentence by filling in missing words in a sentence frame. In order to fill in the missing words students may use words, leave out words, and change words in the original sentences, or they may add new words. TAS was chosen for this project, because it supposedly assesses a student's ability to subordinate in writing, and subordination is one aspect of syntactic elaboration. At the end of the term, the TAS was administered again as a post-test. In addition, during the last week of the term a writing work-sample was collected from the students. The compositions were marked according to three methods outlined by Heaton:

> The impression (holistic) method of marking entails one or more markers awarding a single mark, based on the total impression of the composition as a whole .... Impression marks must be based on impression only, and the whole object is defeated if examiners start to reconsider marks and analyze compositions. The [analytic] method depends on a marking scheme which has been carefully drawn up by the examiner or body of examiners. It consists of an attempt to separate the various features of a composition for scoring purposes .... A third method of scoring compositions is the mechanical accuracy or error-count [error] method. Although this is the most mechanical of all methods of scoring, it is the least valid and is not recommended. (Heaton 1975:136-137).

Each paper was read and scored independently by the author and a colleague. For the holistic scoring a range of 1 to 5 was used. Descriptions for the scale are found in Appendix A. Appendix A also contains the evaluation scheme for the analytical scores. The average of the two readers' scores for each paper for each scoring scheme was recorded as the reported score. Each mistake--mechanical, grammatical, cohesive, coherent, etc.--was counted, all having equal weight.

Various positions abound concerning the nature of composition and its evaluation. Heaton (1975:138-139) states that the ability to write for a particular audience using the most appropriate kind of language is essential for both native-speakers and foreign students. At the advanced level the use of correct registers is an important skill. Furthermore, Heaton claims that the use of appropriate register implies an awareness of both the goal and intended audience.

Harris (1969) values composition tests because they require students to organize their own answers, expressing them in their own words. Composition tests, according to Harris, measure abilities to organize, relate and weigh materials more effectively than do objective tests. Allen and Davies (1977) value composition testing because it provides a natural situation for a student to use what language has been learned--a pragmatic, natural use of language.

While one can find diverse opinions on the nature of composition and its entailments, even wider disagreement prevails about how to evaluate composition and what components should be evaluated. The belief that subjective evaluations are less accurate than the more objective scoring methods has long been held. Heaton (1975:134) lists the following factors that may lead to unreliable subjective scoring: marks may be awarded on (i) what the testee has written; (ii) what the marker believes the student meant by what he wrote; (iii) handwriting and general appearance of what the student has written; and (iv) previous knowledge of the student.

On the contrary, Oller (1979) believes that subjective judgments are indespensable in decisions concerning a writer's expression of intended meaning and how well the intended meaning has been expressed. Research supports Oller's claim. Kaczmarek (1980) found that subjective methods of evaluating compositions work about as well as objective scoring techniques and are strongly correlated with other measures of ESL proficiency having independent claims to validity. In addition Kaczmarek found that subjective teacher judgments have substantial reliability and are strongly correlated with similar judgments by independent raters and with objective scores computed over the same essays. Cooper and Odell (1977) endorse holistic scoring methods, because they consider a discourse as a whole entity and because they offer solutions for placement and exiting grading.

Many objective measures have been used in composition evaluation. Cohen (1980) discusses the use of student errors in scoring student compositions for both testing and testing activities. Witbeck (1976), Metzger (1978), Gorman (1979) and Knapp (1972) also advocate the scoring of errors. For these reasons an error count was included in this study.

One additional objective measure, words per T-unit, was included. Metzger and many native-speaker composition researchers have used this measure in measuring the growth of student writing (cf Kerek, Daiker and Morenberg).

1.3 Findings

The summary statistics for the various measures are found in Table 1; the maximum scores for each measure are enclosed by parentheses. To answer the research questions, a correlation matrix was calculated for all variables; see Table 2.

For this particular subject pool, there was no significant correlation between syntactic enhancement, as measured by TAS (post-TAS) and overall writing ability, as measured by holistic scoring (r=.48). However, the TAS does seem to have predictive validity, because the TAS pretest correlated significantly with the holistic scores obtained at the end of the term (r=.70, p <.01). From Table 2 it can be seen that the TAS has some predictive validity but no concurrent validity. Perhaps the discrepancy can be accounted for by the seemingly inherent lack of reliability of holistic scoring. The correlation of the two independent raters was .58, yielding 34% true overlap between raters.

The pre-TAS correlated significantly with errors (negative), the relevance and fluency atomistic scores. The post-TAS variable correlated significantly with errors (negatively). On the other hand, both the holistic scores and analytical scores correlated significantly with all other variables except words per T-unit. The error variable showed a significant negative correlation with all other variables except words per T-unit, vocabulary and grammar.

The findings of this study are relevant to other studies in terms of the robustness of the atomistic scores. Mullen (1980) found that ratings on vocabulary usage played the heaviest role in determining the overall quality of the composition and that ratings on compositional organization played the least. However, in the current study, the relevance and fluency variables correlated with more variables than did vocabulary.

Evola et al (1980) found a significant though weak correlation between development of overall proficiency and the ability to use grammatical items as hierarchically arranged by first language studies. In this study, the grammar variable correlated significantly with the holistic analytical relevance and fluency scores.

In sum, TAS seems to have predictive validity for eventual (15 weeks) attained writing proficiency, assessed by a holistic marking scheme; however, it does not seem to exhibit any concurrent validity, assessed by holistic and analytical schemes. For this particular study, both the holistic and analytical scoring schemes have similar robustness in terms of significant correlations with other composition variables.

Table 1

Summary Statistics

	Pre-TAS	Post-TAS	Holistic Score	Analytical Score	Errors	Words/ T-Unit
Mean	28.92(45)	33.07(45)	2.51(5)	12.64(25)	23.42	15.82
S.D.	7.91	5.56	.95	2.40	11.86	8.38
	Relevance	Fluency	Vocabulary	Grammar	Mechanics	
Mean	2.92(5)	2.71(5)	2.78(5)	2.14(5)	2.07(5)	
S.D.	.61	.61	.42	.66	.61	

Table 2

Correlation Matrix

	Pre-TAS	Post-TAS	Holistic Score	Analytical Score	Errors	Words/ T-unit
Pre-TAS	1.00	.71 (p < .01)	.70 (p < .01)	.02	-.82 (p < .01)	.07
Post-TAS		1.00	.48	.43	-.71 (p < .01)	.07
Holistic Score			1.00	.95 (p < .01)	-.75 (p < .01)	-.16
Analytical Score				1.00	-.67 (p < .01)	-.01
Errors					1.00	-.14
Words/T-unit						1.00

Table 2

Correlation Matrix

	Relevance	Fluency	Vocabulary	Grammar	Mechanics
Pre-TAS	.77 ($p<.01$)	.58 ($p<.05$)	.38	.26	.36
Post-TAS	.43	.43	.10	.39	.40
Holistic Score	.85 ($p<.01$)	.73 ($p<.01$)	.72 ($p<.01$)	.78 ($p<.01$)	.78 ($p<.01$)
Analytical Score	.86 ($p<.01$)	.87 ($p<.01$)	.59 ($p<.05$)	.85 ($p<.01$)	.85 ($p<.01$)
Errors	-.72 ($p<.01$)	-.67 ($p<.01$)	-.51	-.46	-.52 ($p<.05$)
Words/T-unit	.22	.004	.14	-.15	.07
Relevance	[illegible]00	.76 ($p<.01$)	.52 ($p<.05$)	.59 ($p<.05$)	.62 ($p<.05$)
Fluency		1.00	.34	.68 ($p<.01$)	.67 ($p<.01$)
Vocabulary			1.00	.39	.35
Grammar				1.00	.72 ($p<.01$)
Mechanics					1.00

Bibliography

Allen, J. P. B. and Alan Davies (eds). 1977. *Testing and experimental methods, Vol. 4, The Edinburgh course in applied linguistics*. London: Oxford University Press.

Cohen, Andrew D. 1980. *Testing language ability in the classroom*. Rowley, Mass.: Newbury House Publishers, Inc.

Cooper, C. R. and L. Odell (eds). 1977. *Evaluating writing*. Urbana, Il. National Council of Teachers of English.

Davidson, David M. 1978. *Test of ability to subordinate*. New York: Language Innovations, Inc.

Evola, Jill, Ellen Mamer and Becky Lentz. 1980. "Discrete point versus global scoring for cohesive devices". In John W. Oller, Jr. and Kyle Perkins (eds). *Research in language testing*. Rowley, Mass.: Newbury House Publishers, Inc. 177-183.

Gorman, Thomas P. 1979. "The teaching of composition". In M. Celce-Murcia and L. McIntosh (eds). *An introduction to teaching English as a second or foreign language*. Rowley, Mass.: Newbury House Publishers, Inc. 189-202.

Harris, David P. 1969. *Testing English as a second language*. New York: McGraw-Hill.

Heaton, J. B. 1975. *Writing English language tests*. London: Longman Group Limited.

Kaczmarek, Celeste M. 1980. "Scoring and rating essay tasks" In John W. Oller, Jr. and Kyle Perkins (eds). *Research in language testing*. Rosley, Mass.: Newbury House Publishers, Inc. 151-159.

Kerek, Andrew, Donald A. Daiker and Max Morenberg. 1980. "The effects of intensive sentence combining on the writing ability of college freshman". *Language and Style XIII* 3: 81-90.

Knapp, Donald. 1972. "A focussed, efficient method to relate composition correction to teaching aims". In H. B. Allen and R. N. Campbell (eds). *Teaching English as a second language*. New York: McGraw-Hill. 213-221.

Metzger, E. 1978. "A scheme for measuring growth in college writing". *Journal of Basic Writing 1* 4: 71-81.

Mullen, Karen A. 1980. "Evaluating writing proficiency in ESL". In John W. Oller, Jr. and Kyle Perkins (eds). *Research in language testing*. Rowley, Mass.: Newbury House Publishers, Inc. 160-170.

Oller, John W. Jr. 1979. Language tests at school. London: Longman Group Limited.

Witbeck, Michael C. 1976. "Peer correction procedures for intermediate and advanced ESL composition lessons". TESOL Quarterly 10 3: 321-326.

Appendix A

EVALUATION FORM FOR SCORING COMPOSITIONS

Student's name______________________________

	5	4	3	2	1
Relevance					
Fluency					
Vocabulary					
Grammar					
Mechanics					

RELEVANCE means adherence to the topic, thesis development with supports and overall unity.

FLUENCY refers to readability, organization, and coherence.

VOCABULARY refers to level of word choice, variety of vocabulary, appropriate usage, and idioms.

GRAMMAR refers to cohesive markers, the complexity of sentence structures, and sentence boundaries (e.g. no fragments or run-ons).

MECHANICS includes minor grammar rules (e.g. subject-verb agreement), spelling, and punctuation.

SCALE

5:	excellent	The essay matches the top standard of native writers.
4:	very good	The essay exhibits only a few minor problems.
3:	good	The essay communicates something of the writer's ideas, but with frequent interference; and/or there is clear communication, but with rather simple presentation.
2:	poor	The essay exhibits serious problems; only half or less of the elements and ideas are coming through; the reader has to struggle.
1:	bad	The essay is almost incomprehensible, difficult to process, and/or is excessively simple, containing only the bare elements.

ED217735

THE C-TEST - A MODIFICATION OF THE CLOZE PROCEDURE

Ulrich Raatz and Christine Klein-Braley (University of Duisburg)

1 Introduction

For the language tester the cloze procedure is a well-known but at the same time a somewhat controversial instrument for measuring language proficiency - whatever that may be - both in L1 and L2. This procedure, which perhaps ought to be called rather a method for constructing tests, is commonly regarded as one of Spolsky's (1975, 1981) psycholinguistic-sociolinguistic (or post-modern) measurement devices. It was originally greeted by language testers with great enthusiasm. In recent years the attitude towards cloze tests has become more critical among the majority of researchers. This more cautious approach is a result of the fact that there are a number of unsolved, maybe even insoluble, problems involved in the development, application and interpretation of cloze tests.

In Duisburg we have been developing a modification of the cloze principle which seems to have a number of advantages over the classical cloze test. We have developed several versions in German and in English and have tried them out in a number of groups of L1 and L2 learners of various ages. This paper will describe the results we have obtained so far.

2 Criticism of classical cloze tests

2.1 Introductory remarks

The classical cloze test is a test consisting of a connected passage of prose in which individual words have been removed in a systematic manner. A number between 5 and 10 is chosen and every nth word is deleted. The examinee confronted with such a test is expected to find suitable slot-fillers for the resulting blanks. His responses are scored in one of two ways: either by the exact method, where the examinee receives a point only if his solution is identical with the word the author of the passage used, or by the acceptable method, in which any word which fits the context grammatically, syntactically and semantically is counted as correct. Very high correlations have been reported between these two scoring methods, although there seems to be a tendency in L2 testing to prefer the acceptable method. Cloze tests belong to the integrative methods of testing language skills, and are usually claimed to measure the construct of general language proficiency or pragmatic expectancy grammar. It seems fair to summarize the general consensus of opinion among language testers as being favourable to cloze tests, which are regarded as objective, reliable and valid tests of L2 proficiency. However in recent years a number of studies have appeared criticising various aspects of cloze tests, and attitudes towards them have become more critical.

FL013025

The positive claims made for cloze tests are based on an immense amount of empirical research and the tests themselves are anchored in a firm framework of what it means to know a language. The best summary of this evidence can be found in Oller's Language tests at school (1979). But it seems to us that some of the conclusions that have been drawn from empirical studies have been interpreted too broadly; that positive results emerging from one study have been generalized too quickly to others; that a number of essential questions have not, so far, been thoroughly examined.

In our view the four main problem areas are the following:

1. text selection
2. test construction
3. scoring
4. interpretation.

The evidence offered in support of our criticism is empirical, and emerged from a study conducted between 1977 and 1981 in the University of Duisburg (Klein-Braley 1981). The subjects were undergraduates preparing to become teachers of English, most of whom had already had eight or nine years of English teacning before entering the university. In all, 19 different English cloze tests and two German cloze tests were used, together with a number of other test procedures. The aim of the study was to examine the performance of cloze tests for this population.

2.2 Text selection

In constructing the first cloze tests used in the Duisburg study one of the major headaches was the question of text selection. Unfortunately this question is not even discussed in the majority of published reports. One author who does discuss the problem is Oller. In Language tests at school he states:

> In spite of the natural feeling that just any old text will not do, or that a suitable text must be a carefully chosen and possibly edited text, research has shown that the cloze procedure is probably appropriate to just about any text (1979:364).

He cautions that certain things must be avoided: emotionally-charged texts; texts requiring esoteric or technical knowledge; texts containing statements that students may strongly disagree with. But this still leaves a wide latitude for choice. Are literary texts permissable? How about older texts? Humorous texts, particularly those whose humour is primarily verbal? Texts using direct speech? Culturally charged texts? Stylistically marked texts? Badly written texts?

In view of the remarks made by a number of experts that the teacher is the best possible judge of what is appropriate for his students, 13 texts were selected in Duisburg as a basis for the cloze tests to be developed. The important question was, of course, whether our expert judgment and that of other teachers would agree. Consequently the texts were submitted to a number of colleagues involved in teaching students in Duisburg or similar students in other universities. The colleagues were asked to evaluate the difficulty of the texts and their suitability for use as cloze tests. The judges were far from unanimous about both criteria, and the majority of texts received the full range of judgments (from very easy to very difficult or from very suitable to very unsuitable) on the seven-point scales given.

A number of the judges commented that the determination of text difficulty was extremely taxing. More serious, however, were the criticisms that some texts were unsuitable because they were not sufficiently neutral. Judges felt that some students might have special knowledge which would enable them to perform better on certain texts. In other words, they were concerned with possible test bias as a result of text specifity.

Confirmation of the problems involved in text finding was provided by a study reported at the Brussels Symposium (Klein-Braley and Klein 1980). This study was conducted to examine the performance of a cloze test in the school setting. While all the teachers involved agreed that the text should be suitable in its level of difficulty for their pupils, the test scores showed that this was not, in fact, the case ($x = 8.63$ out of 50 possible points). The pupils claimed that the subject matter was too specialised.

The problems in this area are therefore:

1. Text selection is a subjective procedure;
2. The tests may be unfair to some examinees as a result of test content (text specifity).

2.3 Test construction

One of the main advantages of cloze tests as opposed to discrete-point item tests is the ease with which they can be constructed. The tester need only find his text, decide on a deletion rate and a starting point, blank out every nth word and his test is ready for use. It is generally recommended that tests should have around 50 blanks.

Klein-Braley (1980) has claimed that there seems to be a general assumption that cloze tests are equivalent to each other, independent of the text, the deletion rate and the starting point involved. This is the logical result of the theory underlying cloze tests. As measures of general language proficiency, making use of random deletion of elements, they force the examinee to activate this trait, and the score obtained is a numerical estimate of the efficiency of this internalized grammar. That the tests are viewed as equivalent is shown by the fact that the tester can choose any text, any deletion rate and any starting point. If the tests are not equivalent, then guidelines are needed for choosing texts, deletion rates and starting points, and it is necessary to specify how different choices will affect the resulting test. Only if the tests all measure the same thing can these factors be considered to be irrelevant. If cloze tests do all set out to measure the construct of general language proficiency, it follows further that any two cloze tests administered to the same group ought to have high correlations with each other. However, very few researchers have actually administered more than one cloze test simultaneously. Those who have done so (Porter 1978; J. Oakeshott-Taylor 1979; Klein-Braley 1980) in fact obtained low correlations between the tests.

Alderson's research (1978; 1979) showed very clearly that even if the same text is used, the deletion rate chosen affected the performance of the tests. When he varied the deletion rate he found different reliability coefficients and different validity coefficients for his tests. These findings were replicated in the Duisburg study in a number of empirical investigations. For instance tests using the same value of n but beginning deletion at different points produced a wide range of values for the difficulty indices, the reliability and the validity coefficients.

In addition, the linguistic assumptions underlying cloze tests were examined. Deletion of every nth word is assumed to produce a random sample of the elements in the text, and, by implication, even of the elements of the language. Since the deletions are performed on individual words, each word in the 13 texts was coded into one of ten grammatical categories. Then, in a computer simulation, 45 cloze tests were constructed for each text, beginning at deletion rate 5 and finishing at deletion rate 10. The words deleted were regarded as a sample of the text, and each sample was compared with the original text to discover whether it was statistically significantly different or not. A chi-square significance test was computed. Since 45 samples were taken from each text, not more than 3 samples should have been significantly different at the 5% level. Only four texts met this requirement; the other nine texts produced more non-random samples than would be expected if nth word deletion functioned satisfactorily as a sampling device. For some texts as many as 10 out of the 45 samples were significantly different. This means that the every nth word procedure does not perform adequately as a sampling procedure, and that one of the basic assumptions underlying the classical cloze procedure is invalid. This problem appears insoluble.

In a further investigation it was possible to demonstrate that the difficulty of any cloze test is related to the ratio of structure and content words deleted from the text. Structure words are easier to replace, and consequently any test where more structure words are deleted will be easier than one which has lost more content words. If the cloze procedure produced random samples of the words in a text the ratio of structure and content words would always be more or less the same, but as the computer simulation showed, this assumption is invalid. The correlation between test difficulty and the proportion of structure words deleted for the tests in the study was .30 for acceptable scoring and .46 for exact scoring.

Other results from the Duisburg study point in the same direction. One factor which struck us forcibly was the question of text length. It was necessary to produce a number of tests from the same text, and it became very obvious that if an every 10th word deletion rate is chosen one needs twice as much text (around two typed pages) as for an every 5th word deletion rate. However, despite their length, such tests have a relatively small number of items.

These findings, taken together with those of Carroll, Carton and Wilds (1959), Porter (1978) and Alderson (1978,1979) add up to the conclusion that the qualities of a cloze test are dependent on the text chosen, the starting point and the deletion rate. This means that every single cloze test needs to be examined to determine its individual characteristics. There is no justification for the assumption that every cloze test will perform satisfactorily. In other words, the assertion that every cloze test is automatically reliable and valid must be rejected.

There are therefore five major problems in this area:

1. The cloze construction principle (deletion of every nth word) does not ensure random deletion of the elements in the text;
2. Each of the factors text, deletion rate, starting point affects the quality of the test as assessed by the reliability and validity coefficients;
3. The ratio of structure/content words deleted affects the difficulty of the test;
4. The test constructor cannot tell in advance how his text, his deletion rate and his starting point will affect the characteristics of the test;
5. Tests using the higher deletion rates (8, 9, 10) need extremely long texts if 50 items are to be ensured.

2.4 Scoring

A test is objective when the results obtained are independent of the scorer involved. Objectivity is not a problem when cloze tests are scored using the exact method, as the original text is available as a test key. However there are two disadvantages when exact scoring is used for non-native speakers:

1. Scores are rarely distributed over the entire range of the scale. It is not unusual to find that the highest scorer has obtained only half to two thirds of the points available;
2. Examination of test scripts shows that non-natives are often able to supply appropriate slot-fillers, even though these are not identical to the original word.

It therefore seems questionable whether cloze tests scored by the exact method are appropriate for non-natives. Proficiency in the foreign language is surely demonstrated when an examinee fills a blank in an acceptable manner, even though his solution may differ from the original text. So acceptable scoring is often felt to be more appropriate. However, at this point the problem of objectivity reappears, since the scorer must make judgments regarding acceptability, and it is open to doubt whether different judges would agree in their assessments.

Onli Hanzeli (1977) and Naccarato and Gilmore (1976) have reported investigations of inter-rater reliability. Both studies showed that this was satisfactory, however in both cases the subjects appear to have been beginners or intermediate students. Our research in Duisburg, performed with advanced students, contradicts these findings. Scorers disagreed about the acceptability of a solution offered in no less than 20 out of the 30 items in one test, although individual judges were consistent in their acceptance or rejection of individual items.

Examination of test scripts with high discrepancies between acceptable and exact scoring shows that the very proficienct student is able to find a filler for almost every blank. This probably also explains the discrepancy between our findings and those of other researchers as far as the correlations between the two types of scoring are concerned. For the 19 tests in the Duisburg study the correlations ranged from .45 to .90 with the majority of values in the .6 to .7 region. Another reason is probably that in no case were fewer than four native speaker experts involved in the scoring, and, since the final decision on acceptability was a group decision, the tolerance limits tended to be broader for the group than for any one individual scorer.

The type of scoring method used turned out to be a further feature affecting the performance of the test so far as reliability and validity were concerned. Acceptable scoring produced more reliable tests for students who had never completed a cloze test before. However, after students had received training in completing cloze tests, exact scoring produced the more reliable results.

The problems with scoring are therefore the following:

1. Exact scoring, while objective, is inappropriate for L2 learners unless they have received training in completing cloze tests;
2. Acceptable scoring is not objective;
3. The scoring method used affects the performance of the test;
4. The two methods of scoring are not necessarily equivalent for advanced L2 students.

2.5 Test interpretation

As a norm-oriented test, the classical cloze test is intended to determine the status of an individual relative to others in a specific population. The question of population is a crucial one here: a test developed according to classical test theory is always designed for a specific population and can only be interpreted in terms of the population it is intended to measure.

In much of the research into cloze tests the population has been assumed to be all learners of a second language, whatever their background or native language or previous learning experience. Thus many of the decisive studies have been performed in the USA on multilingual samples intending to study at American universities. Subjects in these studies appear to range from complete beginners to very highly proficient individuals. The results of such studies have been positive. Cloze tests have proved to be extremely reliable and to have high correlations with validation measures used.

The question is whether results derived from such groups can be transferred to other, differently structured groups. The subjects involved in the Duisburg study were not comparable to the subjects in many of the American studies. They were monolingual in their L1. They did not represent the entire range of learners: only advanced students of English were involved. Thus we have a multilingual group on the one hand and a monolingual group on the other hand. Moreover the monolingual group is restricted in range.

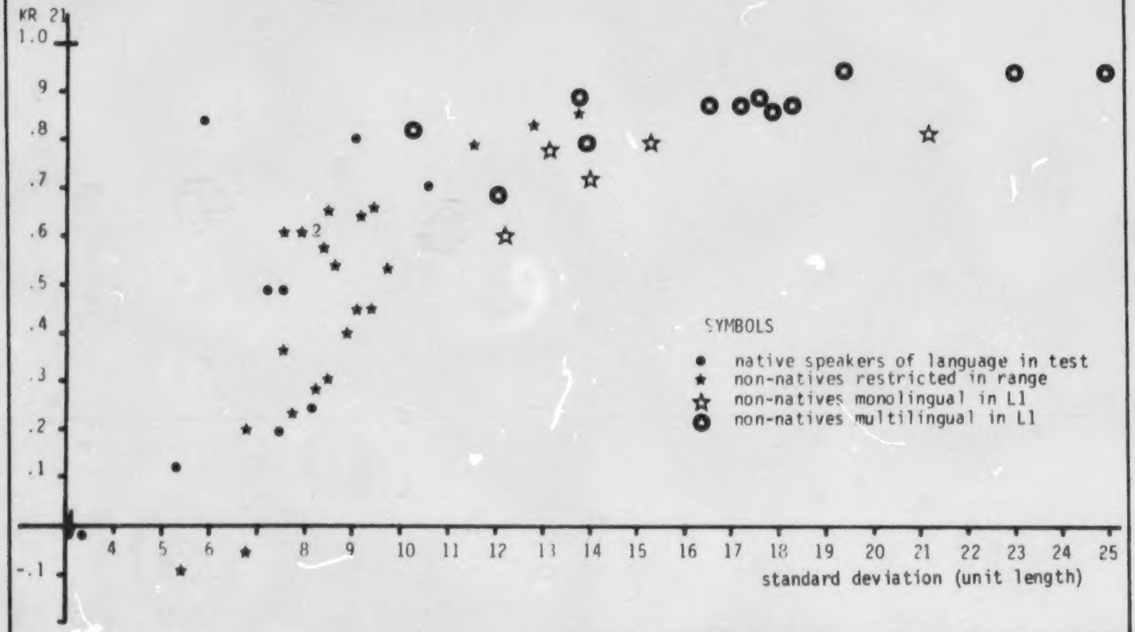
Restriction in range is known to affect the statistics of a test negatively. All correlations are lower: item discrimination indices, reliability coefficients, validity coefficients. This is precisely what emerged in the Duisburg study: the tests failed to reach the .9 reliability level and validity could not be demonstrated. The important question is whether this is an isolated finding. Is the Duisburg group unique or is there some kind of regularity operating here?

The differences between the groups used in the Duisburg study and those in other studies have been described above. We now set out to find other studies using cloze tests with groups similar to ours. In all we analysed data for 55 different groups reported in the literature. In order to make comparisons we had to standardize the results, and so KR 21 coefficients were calculated for all tests from the data given in the reports. All standard deviations were adjusted to apply to unit length tests. This was necessary because in some studies reliability coefficients are not reported at all, and the tests used in different studies have different numbers of items. The results of this investigation are graphed in Figure 1. In Figure 1 four groups with varying degrees of homogeneity have been defined: natives; non-natives restricted in range (as in Duisburg); monolingual non-natives; multilingual non-natives. The Duisburg results are not included in this graph. The small black stars symbolize groups which are similar to the Duisburg groups, and it emerges very clearly that the majority of tests for such groups are unreliable if we take the .9 level as our criterion. There is an obvious curvilinear relationship between the standard deviation of the tests (s/n%) and the reliability coefficients and there is a marked tendency for the standard deviations of the restricted range non-natives to be small.

This shows that the reliability coefficients obtained are dependent on the type of sample involved. However, tests with low reliability coefficients are measuring instruments with a wide margin of error. At the same time the small standard deviations show that only a narrow range of scores were obtained on the tests. Thus these test results cannot be interpreted in terms of individual differences. This phenomenon applies to all the tests symbolized by small black stars in Figure 1. Where were these restricted range non-natives tested? In general at universities in a European country or in the L2 classroom. But given the huge standard error of measurement of these tests, they are of little practical use for the teacher since he cannot legitimately base any decisions on them.

The second problem is the question of validity. What do cloze tests measure? In the American studies high correlations are reported between cloze tests and traditional discrete-point item tests, thus suggesting that cloze tests do indeed test general language proficiency whether this be viewed as an additive or a holistic construct (cf. Farhady 1979). But in the very few studies using cloze tests with restricted range non-natives reporting correlations with criterion measures, these correlations have been much lower. Only medium to low correlations were obtained between the cloze tests used in Duisburg and the DELTA test (cf. Klein-Braley and Lück 1980), a highly reliable psychometric-structuralist test. Similar results are reported by Carroll et al (1959) and by Macha (1979). In the school context Hanzeli (1977) reports correlations between Grade Point Average and

Figure 1 KR-21 plotted against unit-length standard deviation



cloze tests varying between .69 and -.15, and in their study Klein-Braley and Klein (1980) found a correlation between cloze scores and teacher assessment of .08. It seems therefore that the classical cloze test does not succeed in measuring what teachers view as important.

A final question is that of relating cloze test scores to some known level of proficiency in language skills. Virtually all psychometric-structuralist tests make one important assumption: that an educated native speaker of the language would achieve a more or less perfect score on the language test. This seems to us to be an eminently reasonable and sensible assumption. It was, in fact, Carroll's approach to cloze tests in 1959: he intended to interpret the learner's score in relationship to the scores obtained by adult speakers of the language in question. He was, however, thwarted by a peculiarity of the classical cloze test: native speakers do not perform perfectly on these tests. If they did, cloze tests could not be used in L1 testing. A high score on a psychometric-structuralist test, therefore, is a definite indication that the examinee is approaching a level which can be related to native speaker performance. If the native speaker averages 99% on a particular test, then the foreigner who scores 95% is doing pretty well. But how can cloze scores be interpreted? In the Klein-Braley and Klein study (1980) the highest score was 28 out of 50 or 56%. In terms of the group as a whole it was a good score, of course, but what did it tell us about this individual's actual level of performance? The fact that each researcher uses his own self-developed cloze tests compounds this problem even more. Results derived from cloze tests are not comparable to each other, and this makes it very difficult to assess what a score on any one test tells us about the individual.

Thus the problems in the area of test interpretation are the following:

1. Cloze tests are not suitable for restricted range groups since they are not sufficiently reliable and do not produce adequate variance;
2. Classical cloze tests are of little use in the L2 classroom situation since individual test scores cannot be interpreted in terms of individual differences;
3. Empirical studies have shown only weak relationships between cloze test scores and teacher judgments, suggesting that cloze tests do not produce the information the teacher needs;
4. Scores obtained from cloze tests cannot be easily related to native-speaker performance. This makes judgment of a student's relative proficiency on the basis of his cloze score rather difficult.

3 Theoretical considerations leading to the development of the C-test

3.1 Summary of the criticism of the classical cloze test

We have shown that there are major problem involved in the construction and use of classical cloze tests. The major problems can be summarized as follows:

The deletion rate used in the classical cloze procedure is too large. If we demand at least 50 items, then extremely long texts are needed. In addition, the deletions produced are not a representative sample of the different types of words in the passage. This, in its turn, produces the lack of comparability in reliability, validity and difficulty when the same text is used with different starting points and deletion rates.

If only one text is used, this cannot be expected to be a representative sample of the language as a whole. In addition, it can easily happen that examinees with special interests or knowledge have unfair advantages, others unfair disadvantages.

The two different methods of scoring - exact and acceptable - are both problematical. If exact scoring is used cloze tests are too difficult even for competent native speakers. In addition, it is inappropriate for the non-native examinee. The acceptable method is not objective or, if efforts are made to ensure objectivity, it is extremely time-consuming. The high correlations between the two methods of scoring which are often used as evidence for their equivalence are systematically biased because of the large number of items which are common to both scoring systems.

It seems axiomatic that a language test designed to measure the performance of the L2 student should produce virtually perfect scores when administered to the L1 speaker. Linguistically speaking it is difficult to see how the correct solution for a test item can be determined other than by basing notions of correctness on the general consensus of native speakers of the language. This condition is one which, in practice, is not fulfilled by any cloze test.

3.2 Characteristics to be included in a new procedure

We felt that it ought to be possible to retain the theoretical basis underlying the classical cloze procedure - the concept of an internalized grammar which is utilized in all language operations - but to develop a new type of test form which would not suffer from these defects. Our criteria for the new procedure were as follows:

1. It ought to use much shorter texts but produce at least 100 items;
2. There ought to be no problems involved in the choice of deletion rate and starting point;
3. The deletions ought to be an absolutely representative sample of the elements of the text;
4. It should not favour examinees with specialised knowledge;
5. It should use only exact scoring, and thus be entirely objective;
6. Native speakers should obtain virtually perfect scores.

In addition the new procedure should produce reliable and valid tests and be relatively easy to develop.

3.3 The principle underlying the C-test

Our new procedure is a cloze test in which every second word is deleted. However, in order to ensure that solution is possible at all, we leave the first half of the deleted word standing. If the word has an odd number of letters, we delete exactly half the word plus half a letter. If the word consists only of one letter, as in the English words a and I, then this word is ignored in the counting, and half the following word is deleted.

The examinee restores the missing parts of the words. Only entirely correct restorations are counted as right.

3.4 First exploratory investigation with the new principle

In the first pilot study we wanted the answers to three questions:

1. Does the principle function at all?
2. Do competent native speakers achieve near-perfect scores?
3. Do the tests produce a wide range of scores when administered to L2 speakers?

In order to examine these questions we developed a German test form with 103 items and an English form with 138 items. Both test forms were given to 10 English native speakers with German as L2, and to 10 German speakers with English as L2. Our subjects were English lectors, English army officers, German professors and advanced students in the University of Duisburg. The results conformed to our expectations. Both groups of native speakers scored at least 95% on the native language versions. Their results for the L2 tests were distributed over the whole scale. These results encouraged us to continue investigating the new procedure.

3.5 Second exploratory investigation

This investigation was aimed at revealing whether deletion of every second word in a 200 word text - our initial approach - produced a representative sample of all the words in the text. Two students were asked to examine 40 English and 40 German texts, each with 200 words, chosen at random to vary as much as possible so far as style and content were concerned. They coded each text into the categories: nouns, adjectives, verbs, adverbs, prepositions, articles and remainder. Then they deleted every second word and compared the distribution of the deleted elements to that of the elements in the original texts. The results showed that for every one of the 80 texts the deleted samples are representative for the entire text.

3.6 Development of the final form

One problem remained: the question of text specifity. We therefore decided to use not one long text but several short ones which were unconnected with each other. This had two advantages:

1. Even if an examinee has previous knowledge of a text or specialised knowledge in a particular area, his advantage over the other examinees is not so large. This means that the test is fairer;
2. The individual parts can be viewed as super-items, for

which difficulty coefficients, discrimination indices and intercorrelations can be calculated. This means that an item analysis is possible. Furthermore the reliability of the test can be calculated using Cronbach's Alpha.

For the German version we decided to use five parts with 20 items each. For theoretical reasons it seemed sensible to increase the difficulty from section to section. Deletions were not to begin immediately: one or two lines of text were to be left as a lead-in, just as in the classical cloze. However, because either of the two possible cloze tests that can be developed from any one text is a random sample, the starting point is irrelevant. The German version would thus have around 250 words providing 100 items.

In the development of the English version we noticed that English words tend to be shorter on average than German words, and that consequently a smaller quantity of language material was involved. In order to adjust for this phenomenon, the English version was to have 5 texts with 25 items each, and would thus have around 280 words.

The scoring was to be exact; in any case because of the half-words that are left standing there are very few cases in which alternatives are possible.

The tests which were developed according to this principle conform very closely to the criteria listed above. Whether they would be useful in practice was a matter for empirical study.

4. Investigations using the English versions of the C-test

4.1 Development of the English versions

The English versions were to consist of 5 short texts of increasing difficulty, each with 25 blanks, thus giving 125 items. The two target groups envisaged were:

1. The Duisburg placement group. These are students entering the university after approximately 9 years of English at school. They can be considered advanced L2 learners;
2. English primary school pupils in the third class of the junior schools. These are nine and ten-year-old native speakers of English.

The most difficult problem was to find texts of an appropriate level. It seemed unlikely that the same set of texts would perform adequately for both groups. Intuitively it seemed that easier texts would be needed for the primary school pupils than for the Duisburg group. We therefore decided to construct a "junior" test and a more difficult "senior" test. For the junior test we used paragraphs taken from the Ladybird graded readers for three texts and from children's books for two texts. For the senior test the material was taken from English language teaching text books, and all texts were allegedly authentic. Readers were used as sources: from the tenth and the 13th grade, from the beginning and end of a Cambridge First Certificate reader and the fifth from a Cambridge Proficiency Certificate reader.

4.2 Control group experiment

In this experiment using the senior test we were interested in the answers to the following questions:

1. Is there only one possible answer to each of the blanks or are there items where more than one answer is possible?
2. Do proficient native speakers obtain a virtually perfect score on the test?

The test was administered in a V class in a comprehensive school in Leeds. The instructions to the pupils (in this and in all the experiments) ran roughly as follows:

> "You will find 5 short texts in which the second half of every second word is missing. Try to guess what is missing and write it in the blank"

The 21 pupils involved were all pre-O-level students, and all were girls. We obtained the following results:

1. For 7 of the items on the test there are viable alternatives. In particular the deictic elements this/that permit alternative solutions. However in no case is more than one alternative possible;
2. No-one in the group made perfect scores. The highest score was 124 and the lowest 111. The median score was 118 (94%).

While these scores are not perfect, they are extremely high, and since these young people are not yet adult native speakers, we would expect, on replication with adults, to obtain even higher scores. The control group investigation shows that this version of the C-test can be scored objectively and that native speakers obtain very high scores.

4.3 The performance of the L2 version of the C-test (English) for students of English as a foreign language

In April 1981 the freshmen in the English faculty of the University of Duisburg were administered the English senior version of the C-test. We were looking for answers to the following questions:

1. What are the difficulty coefficients and the discrimination indices for each test section and for the entire test for this sample?
2. How high is the reliability of the test?
3. How high is the concurrent validity with a suitable criterion?

The C-test was administered as part of the placement test. The DELTA test, run in the same session, was actually used for placement and served as the validation criterion. The sample consisted of only 27 subjects, 12 male and 15 female. They had had between 5 and 13 years of previous instruction in English. The subjects were all L1 German speakers. DELTA contains subtests for grammar (GRA1, GRA2, GRA3) and vocabulary (VOC1, VOC3) and a dictation (DICT). A further test administered in this session was another variation on the cloze procedure (ICLOZE).

The analysis of all the available data gave the following results:

The range of scores on the C-test was from 87 to 117 with a mean of $\bar{x} = 104.67$ and a standard deviation of $s = 9.29$. The average difficulty of the test was 84%. The test was therefore much too easy for these subjects. The difficulty coefficients (P) and item discrimination indices (r_{jt}) are shown in Table 1.

Table 1

Difficulty and discrimination indices for the five test parts

Part	P	r_{jt}
1	83.70	.52
2	94.66	.40
3	80.84	.78
4	78.07	.59
5	81.84	.78
Mean intercorrelation of 5 parts = .46		

These values are not satisfactory. In particular the test is far too easy. Considerable improvements are necessary here.

The reliability of the test was estimated using Cronbach's coefficient Alpha, the generalized version of Kuder-Richardson 20. The result was a reliability coefficient of $r_{tt} = .81$. In view of the less-than-perfect difficulty level of the test this is extraordinarily high. It implies that a test of the appropriate difficulty for these subjects would probably produce a satisfactory reliability. The concurrent validity with the DELTA score as criterion was $r_{tc} = .72$.

In order to discover what the C-test actually measures a factor analysis was performed. Table 2 shows the correlations of the C-test with the individual DELTA subtests and the factor loadings on the first factor which explains 62% of the variance (common factor model, varimax rotation).

Table 2

Intercorrelations of C-test with DELTA subtest and factor loadings on first factor

DELTA Subtest	C-test correlations	Loading Factor 1
GRA1	.56	.80
VOC3	.63	.92
GRA2	.36	.52
GRA3	.52	.83
VOC1	.50	.80
DICT	.60	.57
ICLOZE	.61	.80
C-TEST		.72

We can summarize these results by stating that despite certain unsatisfactory aspects of the test it has a surprisingly high reliability and validity. It loads together with the other tests administered on a factor which can probably be best described as a "general language factor". We assume that a test of the appropriate level of difficulty for this sample would have entirely satisfactory reliability and validity coefficients.

4.4 Investigation of the performance of the English junior C-test for third year juniors

The next experiment examined the performance of the test for a sample of English schoolchildren in the third year of their junior school. All were native speakers of English. The questions to be answered were:

1. What are the difficulty coefficients and the discrimination indices for each test section and for the entire test for this sample?
2. How high is the reliability of the test?
3. How high is the validity of the test?

The investigation was performed in July 1981 using children taken from two classes of a school in Barnsley. The sample consisted of N = 44 pupils with approximately equal numbers of boys and girls. Their ages were between 9 and 11 with a median of 9.95.

The teacher demonstrated the principle behind the test using the trial items provided. No time limit was set. Since the English primary school does not have a fixed grading system the teacher was asked to assess the pupils' ability in English using a scale ranging from 1 to 6. As in the German system 1 was high and 6 was low. This provided the validation criterion.

Analysis of the data provided the following results:

An analysis of variance showed that there were no significant differences between the mean test scores of the two classes. It was therefore permissible to pool the data. The total test scores ranged from 1 to 93 points. The mean score was $\bar{x} = 55.9$, the overall difficulty 44.8% and the standard deviation was $s = 25.09$. The test has a medium level of difficulty. The difficulty coefficients (P) and the item discrimination indices (r_{jt}) for the five parts are shown in Table 3.

Table 3

Difficulty and discrimination indices for the five test parts

Part	P	r_{jt}
1	71.36	.70
2	60.88	.80
3	33.08	.81
4	27.09	.78
5	30.36	.77

Mean intercorrelation of 5 parts = .67

Improvements could be made by substituting a slightly easier text for text 3 and by interchanging parts 4 and 5.

Again the reliability of the tests was calculated using Alpha. For the entire sample the value was $r_{tt} = .91$.

Concurrent validity was calculated using the correlation between the assessments made by the teacher and the test scores. The assessments ranged from 1 to 6 with a mean value of 3.37 and a standard deviation of 1.6. For the entire group the correlation between test and criterion turned out to be $r_{tc} = -.73$.

The test appears to be measuring the same thing that the teacher uses as a yardstick for his judgments of skills in the English language. Thus it would appear that what is measured by the C-test is a global skill.

We can sum up the implications of these results by stating that the C-test has shown itself in these pilot studies to be an extremely promising instrument. The major unresolved problem appears to be the question of the preliminary assessment of text difficulty by the test constructor.

5 Investigations using the German version of the C-test

5.1 Development of the German version

The German versions were to consist of 5 short texts with increasing difficulty, each with 20 blanks, thus giving 100 items. They were to be used with German-speaking pupils in the third grade (nine-year-olds). We were particularly interested in this group because of the problem of dyslexia. However these versions were also to be used with young people and adults learning German as a foreign language at a language school whose previous knowledge of the language might be very slight. For the pilot study it was decided to construct only one test to be used with both groups.

The texts finally selected were found in a rather unsystematic manner. A student searched through German language readers for the first to the fifth grades, looking for suitable texts. Her selection had to be made on an intuitive basis. Finally she narrowed the search down to three 50 to 60 word texts for each grade level. Then we selected from the three texts the one we considered most suitable. Our criterion, the appropriateness of the texts for both the intended groups, was again purely intuitive. It was decided to arrange the texts in order in the test, beginning with the first grade text and ending with the fifth grade text.

While it seemed that we might be on the right track with the German schoolchildren, there was no evidence to tell us whether the same test would be suitable for the foreign students. So we were particularly interested to see what the empirical investigations would reveal here.

5.2 Control group investigation

In this investigation we were interested in the answers to two questions:

1. Is there only one possible answer to each of the blanks?
2. Can proficient native speakers obtain a virtually perfect score on the test?

In summer 1981 we administered the German version to the students attending a seminar in the University of Duisburg. We felt that these undergraduates, the majority of whom were language students, could be considered to be proficient native speakers.

The sample consisted of N=54 students, 23 males and 31 females, aged between 20 and 34 years with a median of 22.6. The test was administered without a time limit. The students took between five and twelve minutes to complete the test. We obtained the following results:

1. For 98 of the blanks there were no alternative solutions. In two blanks there were two possibilities: both the singular and the plural of the mutilated word. Consequently we decided that in future both solutions would be counted as correct;
2. Only 26% of the group made perfect scores. 22% made one mistake, 52% made more than one mistake. The median score was 98.4 out of 100 possible points. This seemed adequate to support our claims.

The control group investigation therefore shows that this version of the C-test satisfies our demands for objective scoring. It also shows that the educated native speaker can obtain virtually perfect scores.

5.3 The performance of the German version of the C-test for students of German as a foreign language

The next step was to examine the performance of the C-test as a test of German as a foreign language for L2 learners. In this investigation we were looking for answers to the following questions:

1. What are the difficulty coefficients and the discrimination indices for each test section and for the entire test for this sample?
2. How high is the reliability of the test?
3. How high is the concurrent validity with a suitable criterion?

This investigation was performed in April 1981 in the Eurozentrum in Cologne. This is one of 25 Eurocentres or Davies's Schools in Europe. These are situated in five different language areas in eight different countries, and students from all over the world are instructed in the language of the country concerned. When students enter the Eurozentrum they take a placement test, and on the basis of their total test score they are assigned to appropriate courses.

The C-test was run at the beginning of the courses one day after the official placement test had been run. The sample consisted of 99 students, 96 of whom took the C-test. 48 subjects were male, 40 female, five did not specify their sex on the cover sheet. Their ages varied from 17 to 57 years with a median age of 21.6 years. They came from 17 different countries all over the world. All subjects had at least a rudimentary knowledge of German.

The C-test was administered to groups of 12 to 16 subjects by the Eurozentrum teachers. No time limit was set; each subject worked until he/she had finished. The time needed for the completion of the test varied between 10 and 42 minutes.

As the criterion for concurrent validity the Eurozentrum placement test was available. This test contains subtests for listening comprehension, grammar, expression and reading comprehension. The analysis of all the available data gave the following results:

The range of scores on the C-test was from 11 to 89 points. The mean score, and consequently the overall difficulty of the test, was $\bar{x} = 51.2$, the standard deviation $s = 17.4$. The difficulty coefficients and item discrimination indices for the five test parts are shown in Table 4.

Table 4

Difficulty and discrimination indices for the five test parts

Part	P	r_{jt}
1	78.9	.74
2	46.6	.69
3	61.5	.88
4	40.3	.85
5	28.7	.83
Mean intercorrelation of five parts = .70		

The difficulty and discrimination indices of the test parts are ideal. The only improvement that needs making is the exchange of parts 2 and 3, so as to ensure a continuous increase in difficulty.

The reliability of the test was estimated using Cronbach's Alpha. The result was a reliability coefficient of .92. This is extremely high when one considers that a preliminary item analysis had not been performed.

The criterion scores, the results derived from the placement test, ranged from 20 to 93 with a mean value of $\bar{x} = 57.1$ and a standard deviation of $s = 17.0$. The correlation between the C-test scores and the criterion scores, the concurrent validity of the test, was $r_{tc} = .82$.

In order to discover what the C-test actually measures the correlations between the four subtests of the placement test and the total C-test score were calculated. Then a factor analysis was performed. The results are shown in Table 5, which also shows the loadings for each of the tests on Factor 1 which explains 71% of the variance.

Table 5

Intercorrelations of C-test with Eurozentrum subtests and factor loadings on first factor

Eurozentrum subtest	C-test correlations	Loading Factor 1
Listening comprehension	.75	.85
Grammar	.70	.76
Expression	.71	.82
Reading comprehension	.58	.66
C-test		.88

We can summarize these results by stating that the C-test performs optimally for this sample; it has ideal statistical qualities; it measures the same thing as the placement test. The measured one-dimensional construct can best be described as "general proficiency in the foreign language".

5.4 Investigation of the performance of the German version of the C-test with German schoolchildren in the third grade.

The next experiment examined the performance of the test for a sample of German schoolchildren in the third grade - the group the test was originally intended for. This was an L1 group. The questions to be answered were:

1. What are the difficulty coefficients and the discrimination indices for each test section and for the entire test for this sample?
2. How high is the reliability of the test?
3. How high is the concurrent validity of the test using school grades in German as the criterion?
4. Is there a connection between the test score and dyslexia?

This investigation was performed in May 1981 using pupils taken from seven third grades in schools in Duisburg. The sample consisted of N = 130 pupils, with approximately equal numbers of boys and girls. Their ages were between 8 and 12 with a median of 9.1 years. 120 pupils were Germans, 3 were Yugoslavs, 2 were Turks. The other five came from five different countries. 7 children, or 5.4%, had been diagnosed as dyslexic. The test was administered in each of the seven classes. A time limit of 30 minutes was set. The teachers provided us with the grades in German - a global assessment. These grades vary from 1 (excellent) to 5 (very poor).

Two analyses of variance showed that there were no significant differences between the mean test scores and the mean school grades for any of the classes. It was therefore possible to pool the data. Analysis of the data gave the following results:

The total scores of the C-test ranged from 2 to 93 points. The mean score, and at the same time the overall difficulty of the test was $\bar{x} = 60.3$, the standard deviation $s = 17.5$. The difficulty coefficients and item discrimination indices for the five parts are shown in Table 6.

Table 6

Difficulty and discrimination indices for the five test parts

Part	p	r_{jt}
1	86.9	.62
2	67.9	.71
3	70.0	.73
4	44.8	.69
5	34.0	.61
Mean intercorrelation of five parts = .55		

These values are extremely satisfactory. Again, however, parts two and three need to be exchanged.

Again the reliability of the test was calculated using Cronbach's Alpha. For the entire sample the value was $r_{tt} = .86$.

Concurrent validity was calculated using the correlation between the school grades in German and the test results. For the seven dyslexics a grade of 5 was inserted - dyslexics do not receive a grade in German. The grades ranged from 1 to 5 with a mean value of $\bar{x} = 2.8$ and a standard deviation of $s = 1.0$. The correlation between test and criterion proved to be $r_{tc} = -.80$. Thus the test obviously measures the same aspects of the performance in the mother tongue (German) that teachers use as a basis for their assessment of their pupils' levels of attainment.

In Germany dyslexia is diagnosed by administering a test of orthography and an intelligence test to those children with unsatisfactory grades in German. Those pupils whose IQ is higher than a certain value, but whose orthography scores are lower than a certain value, are certified as suffering from dyslexia, and they are given remedial teaching in the third and fourth grades. They are not given an assessment grade in German. Incidentally, this method of diagnosis is relatively controversial.

Our seven dyslexics had a mean test score of $\bar{x} = 33.0$. The non-dyslexics had a mean test score of $\bar{x} = 61.9$. This difference is highly significant. Furthermore the dyslexics are almost all to be found at the very bottom of the distribution of test scores.

This result is a further indication of the validity of the C-test since it is generally assumed that dyslexia is a global weakness in the command of the native language. It also indicates that it may be possible to use C-tests for the diagnosis of dyslexia. We are examining this question at the moment.

We can therefore sum up the results of this experiment as showing that the C-test has proved to have extremely satisfactory test qualities for L1 learners of German as well.

6 Future perspectives

The results we have obtained up to now using the C-test are very promising. Nevertheless we feel that it is still too early to recommend large-scale general use since there are still a number of questions to be answered.

6.1 Criticism of the research performed so far

In empirical research a constant, if unwelcome, companion is error. It is always possible that results obtained in an investigation are dependent rather on certain characteristics of the samples involved than on the factors interesting the investigator. As was pointed out earlier, the wide-reaching claims that classical cloze tests are reliable and valid tests for every conceivable group are based mainly on results derived from heterogeneous multilingual groups, and our own research has shown that they fail to reach the necessary reliability and validity for monolingual groups. While we have controlled for this source of error by examining the performance of the C-tests in both types of group, there may well be other factors operating to produce our positive results which we have not yet identified. We therefore intend to replicate much of the research using larger groups.

One of the main difficulties is with the selection of texts. So far we have based our selection mainly on other people's judgments of text difficulty by using readers and other materials specifically selected for certain target groups. In general these judgments have been confirmed by the rank ordering of the text (item) difficulty coefficients. But this approach proved unsuccessful for the English senior C-test used with the Duisburg advanced L2 students.

We need to examine more texts and their performance in different groups. A number of methods are available for determining text difficulty in advance, but these are not unproblematical, and in any case the majority are intended for use with L1 speakers. How far they apply to L2 materials is unknown, and it may well be that any given text has a number of difficulty levels, depending on the L1 spoken by the L2 learner. In addition it may be necessary to control for vocabulary if L2 C-tests are to be developed for the early and intermediate stages of L2 learning. Ultimately we would aim either at being able to specify criteria for selecting texts of an appropriate level of difficulty, or, if this turns out to be impossible, to offer a selection of texts which have been used in a variety of different groups so that their approximate difficulty has been empirically determined.

It would also be interesting to discover whether the C-principle is language-bound. Does it work for languages other than English and German? It might, for instance, be necessary to modify the principle for languages where case markers and other grammatical components are not, as in German or to a certain extent in English, found at the ends of words.

6.2 What does the test measure and where could it be used?

Our claims for test validity are based both on theoretical considerations and empirical findings.

The theoretical basis for the C-test is identical to that for the classical cloze test. Following Oller and Spolsky we posit an internalized language processing mechanism which can be activated by presenting the examinee with a mutilated text and asking him/her to restore the missing elements. The C-test is however more in accord with sampling theory since the deletions in the text are, in fact, random, and it has a stronger relationship to "what it means to know a language" since the test scores can be related to the virtually perfect scores obtained by proficient adult native speakers of the language.

We have been able to show empirically that the results obtained on the C-tests have high correlations with the two criteria which were available to us: teacher judgments and other language tests. The advantage of the C-test over either of these criteria lies quite simply in its economy. Teacher judgments must be based on long-term interaction with the pupil and on acquaintance with a large number of samples of his performance. Discrete-point item tests demand high investments in terms of time and money in their development phase before they can legitimately be used.

This does not mean, however, that C-tests need no development phase and no statistical analysis. On the contrary it is essential that every C-test developed at this point of time should be very carefully examined. There is a strong temptation to make similar assumptions to those operating with classical cloze tests and we are worried that our positive results combined with the simplicity of test construction using the C-principle might encourage teachers to develop and use such tests without examining whether the tests are, in fact, performing as they ought to (1). At the moment, however, not enough is known about the C-test to make general use legitimate. Any C-test developed should be used initially only for experimental purposes, and its difficulty, discrimination, reliability and validity need to be carefully examined.

One fascinating aspect of the C-test research has been the extraordinarily high validity coefficients obtained when the tests were correlated with teacher judgments. The discrete-point item tests which have been used up to now for such areas as orthography, reading comprehension, vocabulary, grammar etc. have, in general, produced very much lower correlations with teacher judgments. These coefficients have always been claimed to result from the unsatisfactory statistical properties of the school grades: their low objectivity and reliability. Our results seem to indicate that, in fact, the grades have satisfactory statistical properties, but that they measure something qualitatively different from the discrete-point item tests, something more global. Our results therefore seem to be a vindication of the teacher judgments.

Given that further research into C-tests confirms the positive results which have been obtained so far, where could the test be used? We see the following possibilities:

1. It could be used at the end of the school year to check on teacher judgments in L1 at the primary level and in foreign language teaching;
2. A teacher taking over an unknown class or group could use a C-test to give him a global idea of the general level of achievement and of the ranking of individuals;
3. It can be used for selection in cases where selection is necessary, although such procedures are always accompanied by both practical and ethical problems;
4. We were able to show in two experiments that C-tests could be useful as placement tests. A number of investigations into this question are already in progress;
5. C-tests could be used for the diagnosis of learning disturbances in both L1 and L2. Investigations into the use of C-tests for the diagnosis of dyslexia are being carried out in Duisburg;
6. It is difficult to develop tests suitable for the measurement of the level of attainment in German for the children of immigrant workers in the Federal Republic of Germany. A pilot experiment using two classes of Turkish pupils in German secondary schools has produced promising results.

6.3 Conclusions

This has been a report of work in progress, not of work completed and a number of further investigations are in various stages of readiness, others are planned (2). The idea for the C-principle evolved in one of our many discussions while the work on classical cloze tests was being performed, and resulted from the observation, made by Raatz, that nth word deletion was not in conformity with random noise as defined by information theory and that the C-principle was a better approximation. We later discovered that we had merely reinvented the wheel, since experiments with syllable deletion were one of the recommendations made as early as 1959 by Carrol, Carton and Wilds. The C-test started life as a plaything for two incurable test-freaks; in the course of our investigations, however, we have come to believe that it may turn out to be a useful instrument in the language testing cornucopia. The question certainly seems worthy of further investigation (3).

Notes

1 It would be a pity if our remarks discouraged anyone who would like to experiment with C-tests but who does not have the technical facilities available. We are therefore very willing to perform data analysis for anyone in this situation. We would also welcome reports from other researchers experimenting with C-tests or other variations on cloze procedure. Contact either of the authors at the University of Duisburg, Lotharstr. 65, 41 Duisburg, Federal Republic of Germany.

2 In the interval between the Symposium and publication of the proceedings further (confirmatory) data has become available. Authors' addresses as above.

3 We would like to take this opportunity of thanking those people in the various institutions involved who helped us by running and taking the tests. Our thanks are also due to a number of students in Duisburg who performed certain parts of the research. In particular we are grateful to Margit Backhaus, our student assistant, who scored all the tests and prepared the data for computation.

Appendix

Examples of C-tests

1 German C-test

Es war einmal ein kleiner Indianerjunge, der hieß Roter Fuchs. Er hatte ein Kanu, um zu angeln. Aber das Ka_______ war i_______ zu kl_______. Roter Fu_______ lief z_______ seinem Va_______ und sa_______: "Ich bra_______ ein vi_______ größeres Ka_______. Baust d_______ mir ei_______?" "Wie gr_______ soll e_______ denn se_______? fragte d_______ Vater. "I_______ möchte d_______ größte Ka_______ der We_______!" rief Roter Fuchs und breitete die Arme aus.

Text 1

2 English C-test (junior version)

Once upon a time, there was a little girl who lived with her mother, who was a widow. They we______ so po______ that o______ day th______ had not______ left t______ eat. T______ little gi______ went o______ into t______ woods t______ play. S______ was s______ hungry th______ she be______ to c______. An o______ woman ca______ up t______ her. "W______ are y______ crying, m______ child?" s______ asked. "Bec______ I am s______ hungry," said the little girl. "Then you shall be hungry no more," said the old woman.

Text 1

3 English C-test (senior version)

The evening of October 30 1938 was just like any other quiet Sunday night to most of the people of America. Many fami______ were a______ home rea______ the pap______ or conte______ listening t______ the ra______. There we______ two prog______ that ni______ which attr______ large audi______. One w______ the pl______ produced b______ Orson Welles. T______ listeners prep______ themselves f______ an ho______ of comfo______ excitement, b______, after t______ opening announ______, the pl______ did n______ start. Instead there was dance music.

Text 3

Bibliography

Alderson, J.V. 1978. The effect of certain methodological variables on cloze test performance and its implication for the use of the cloze procedure in EFL testing. Paper presented at the 5th International Congress of Applied Linguistics, Montreal, Canada.

- 1979. "The cloze procedure and proficiency in English as a foreign language". TESOL Quarterly 13: 219-226

Carroll, J.B., Carton, A.S., and C. Wilds. 1959. An investigation of cloze items in the measurement of achievement in foreign languages. College Entrance Examination Board Research and Development Report. Cambridge, Mass.: Laboratory for Research in Instruction, Harvard University.

Farhady, H. 1979. "The disjunctive fallacy between discrete-point tests and integrative tests". TESOL Quarterly 13: 348-357.

Hanzeli, V. 1977. "The effectiveness of cloze tests in measuring the competence of students of French in an academic setting". French Review 50: 865-874.

Klein-Braley, C. 1980. A cloze is a cloze is a question. Paper presented at the Second International Language Testing Symposium of the IUS, Darmstadt, West Germany. Forthcoming in: Klein-Braley, C. and D.K. Stevenson. 1982. Practice and problems in language testing 2. Frankfurt etc.: Verlag Peter D. Lang.

- 1981. Empirical investigations of cloze tests. An examination of the validity of cloze tests as tests of general language proficiency in English for German university students. Dissertation, University of Duisburg.

- and W. Klein. 1980. Cloze tests at school. Paper presented at the Third International Language Testing Symposium of the IUS, Brussels, Belgium. Forthcoming in: Culhane, T. and M. Lutjeharms. 1982. Practice and problems in language testing 3. Colchester: University of Essex.

- and H.E.Lück. 1980."The development of the Duisburg English Language Test for Advanced Students (DELTA)". In: Grotjahn, R., and A. Hopkins (eds.). 1980. Research on language learning and teaching. Quantitative Linguistics 6. Bochum: Verlag Dr. N. Brockmeyer. 85-126.

Macha, D. 1979. "Reading comprehension of non-native students in English composition at the freshman level. TESOL Quarterly 13: 425-427.

Naccarato, R.W., and G.M. Gilmore. 1976. The application of generalizability theory to a college level French placement test. Paper presented at the Fourth Annual Pacific Northwest Educational Research and Evaluation Conference, Seattle, Washington.

Oakeshott-Taylor, J. 1979. "Cloze procedure and foreign language listening skills". IRAL 17:150-158.

Oller, J.W. Jr. 1979. Language tests at school. London: Longman.

Porter, D. 1978."Cloze procedure and equivalence". Language Learning 28: 333-340.

Spolsky, B. 1978. "Linguists and language testers". In: Spolsky, B. (ed). 1978. Approaches to language testing. Advances in language testing series 2. Arlington, Va.: Center for Applied Linguistics. v-x.

- 1981. "Some ethical questions about language testing". In: Klein-Braley, C. and D.K. Stevenson (eds). 1981. Practice and problems in language testing 1. Frankfurt etc.: Verlag Peter D. Lang.

ED217756

A REVIEW OF ISSUES RAISED IN THE PRODUCTION OF THE ENGLISH LANGUAGE TESTING SERVICE

Ian Seaton (The British Council)

1 Introduction

This is a brief account of some of the issues raised during the five years of design, development and now operation of the English Language Testing Service (ELTS). As such it touches on problems central to the whole practice of language testing but can only indicate in general terms how the teams working on ELTS attempted solutions. Most of these solutions are necessarily a compromise between what is known of language and language testing and the need of an organisation like The British Council for an operational system. Therefore, while many solutions draw on standard theory and practice in testing, certain others require - and have provision for, careful validation confirmation and refinement.

1.1 ELTS is now the main way that The British Council assesses the language adequacy of overseas students who apply for study or training at tertiary level in Britain. It was developed jointly by The British Council and the University of Cambridge Local Examinations Syndicate from 1975 to 1980, when it was first implemented. In its first year it was used in some forty countries with 4,000 candidates taking it; now in its second year there are test centres in some eighty countries with approximately 8,000 candidates taking it. ELTS is made up of five sub-tests: two General tests of reading and listening comprehension and three Modular tests of language study skills in the modes of reading, writing and speaking. The Modular tests are in six versions using texts specific to the subject areas of life sciences, medicine, physical sciences, technology, social studies and 'general academic'.

1.2 The production of what could therefore be labelled an ESP proficiency test battery inevitably involved a large team, never less than three and at times more than twenty. Members were drawn from The British Council, Test Development and Research Unit at Cambridge and outside. The work divided into four stages: Design, Development, Operation and Validation.

2 Design

Work in this first stage covered the areas set out below:

general characteristics	- population to be tested, - information on language ability needed, - function of score reporting,
specifications	- construct of language performance and levels of adequacy, - content of texts and 'items',

FL013026

operational features - different demands from the 'chain' of users,
- resources (knowledge, time, money) needed to design and maintain the Service.

2.1 Specifying content was a major issue in this stage. The design team knew what information the new tests were expected to provide. Broadly speaking they were following J.B. Carroll's 1961 description of a proficiency test:

> "An ideal EL proficiency test should make it possible to differentiate, to the greatest possible extent, levels of performance in those dimensions of performance which are relevant to the kinds of situations in which the examinees will find themselves after being selected on the basis of the test."

In other words, the tests had to account for more than the formal language system - however described - and try to account for language in use in academic settings. This in turn meant relegating standard grammars and the experience of 'test-writers' to secondary sources for the selection of content and trying through survey, description and specification to select the most relevant language skills and activities for the population. The problem of how to organise and control such procedures had then to be solved. Clearly the large number of people involved in surveying and describing, often mixing data collection with observation and then introspection, would need at least a common language of description and an agreed framework of categorisation. Yet there was little by way of research available in this field to act as signposts for the specifying teams. One particular procedure was, however, immediately available and was eventually used to control the Specifications, the communicative needs processor set out in Communicative Syllabus Design (J.L. Munby, Cambridge University Press, 1978). This had the advantage of having been used already by most of the members of the teams in designing syllabuses and teaching materials and while some had reservations about its use for these purposes, they felt it could be better suited to the inevitably more parsimonious nature of language tests. The choice of this procedure did allow the six teams to survey and to select broad language activities and more detailed language skills in a way that allowed comparison and consensus in the next Development stage. It allowed also each language item to be linked, if only within the terms of reference of the framework, to a set of described communicative activities; this has important implications for the validation stage.

2.2 The 'chain' of users was another important issue. The test would have a particular meaning for each of a series of users of the tests who are involved in the whole process of getting an overseas student to a British university. The 'chain' includes the student him/her self, his/her university or employer, the sponsoring agency (British government or International), the specialist advisers, the British Council office in country, the ELTS centre, ELT schools and centres in Britain, the academic registrars of universities and the faculties or departments themselves. The problem here is that each user has a different expectation of the tests. Those involved in language teaching want as detailed information as possible about first, the content of the test, the 'shadow syllabus', and second, what are the particular strengths and weaknesses of 'their' candidates as reported by the tests' scores. Those involved in taking administrative decisions may only want to know an overall score to which they can assign some general meaning and which they can rely on to be secure and consistent. The solution adopted was to report all sub-test scores in the form of a profile, to link that profile to a set of performance descriptions on a scale of one to nine and to provide initial guidance by profile interpretation in terms of levels of adequacy broadly defined for types of courses and predictions of tuition time needed to raise the 'average' learner under 'average' conditions to the appropriate level. It is obvious that this confronts a host of important issues in language testing and teaching but it is equally obvious that the chain of users already requires such information; information that previous systems were not designed to deliver in a coherent form. The decision was taken therefore to link scores to criterion descriptions and to provide initial criterion meaning to the profile by using parameter six of the communicative needs processor, target level-dimensions and conditions. Rather than allow the first few years of operation to accrete criterion meanings of language adequacy to the scores, it was planned both in pre-testing and in the overall validation to provide an initial meaning which could be further and further substantiated and refined. The solution, therefore, not only derives from the application of particular procedures to idealise levels of adequacy but pre-testing on a small longtitudinal sample as well as setting up two units in The British Council to monitor the main validation.

3 Development

Work in the second stage covered:

overall shape	- mix of general and modular sub-tests, - relations between the sub-tests, - band descriptions and profiles,
test realisation	- item types, - control of item writing,
pre-testing	- sampling, - analyses used, - revision.

3.1 Editing the specifications provided by the six teams was a major issue in this stage. While the use of a common framework and common meta-language was a definite advantage, the use of the communicative needs processor as the survey and specifying procedure resulted in over delicate and detailed specifications. While the teams had a prescribed procedure to follow, there was less prescription in their methodology of survey, in particular the mix of data collection and introspection/value judgement. The problem therefore was to reduce the wealth of specification down to a selection for a maximum 3 hour series tests, to ensure that selection represented a common core of language activities and skills and to distinguish which part of this core rested on some kind of objective observation or collection of data and which on the subjective - but valid - view of more than one team of test designers. The solution adopted was that of an editorial group scrutinising each specification and calling in, if necessary, each team to explain, justify with a view to producing a robust set of specifications that rested on the firmest foundations that each team had provided.

3.2 Realising the tests and selecting item types was another main issue. Put simply, specifications describing multi-mode activities and authentic language communication in specified settings cut across the usual boundaries of 'listening, speaking, reading, writing'. Available item types, particularly those objectively scored, did not appear to be able to handle even the most discretely defined skills. For example such low-order skills as in categories one to ten in the inventory contained in Communicative Syllabus Design are nevertheless integrated into performance that depends on features like sequence and purpose. Even so-called integrative item types such as the essay (or dictation or cloze?) seemed inadequate by being too separate from the groups of language study skills specified. However the available item types did have known characteristics and did allow the standard procedures of norm-referencing to be used to establish the reliability of the tests. Moreover, many of the users would expect tests to be made up of such items as guarantors of stability or to satisfy face-acceptance, while test centres would require ease and consistency in scoring the tests. It was therefore decided to realise three of the sub-tests in multiple-choice items and two in integrated tasks that would be directly assessed. Each multiple-choice item would relate to a specified micro or enabling skill; this would allow the editing team to check on the language realisation and later, possibly, the validation studies could see to what extent skills that were expected to cluster together or be dependent on each other actually 'worked' in that way for all, or categories of, candidates. The task items in the written and spoken sub-tests would provide an immediate meaning to a candidate's performance and could enable, if the relation of the sub-test scores was assumed to be linear and equally accumulative, an interim meaning to be established in pre-testing for the whole profile. The compromise on item types could give the system as a whole the necessary 'anchor' of reliability - through pre-testing and monitoring - as well as providing data, over the years, to begin to clarify how different item types performed.

3.3 The tests were pre-tested first in Britain on a sample of some 950 students in 34 centres, edited and then pre-tested overseas on 603 students in 19 countries before being set in their final form. The editing after each pre-test presented a third major issue in this stage. For example, the task based sub-tests are directly and subjectively assessed from band descriptions provided to describe performance at each level. Responses from assessors showed that some preferred the specified features that described each level to be summarised in a criterion 'thumb-nail' sketch of performance. Others preferred to have each feature separately described. The editing team chose a middle way which would be amenable to development in either direction depending on the feedback in monitoring the Service's actual operation. The main problem with the objectively scored tests was how to reconcile the description of an item provided by the standard analyses, index of discrimination, facility value, etc., with the criterion description of the item in terms of it representing this micro-skill at that level of difficulty in that cluster of enabling skills. In other words there had to be a trade off for those items which were deemed unacceptable in norm-referenced terms yet which represented an important part of the criterion skills specified. In practice this meant rewrite rather than straightforward rejection to arrive at the 'best' fifty items. It was recognised that although an item might appear overtly to be testing a particular skill, covertly it might be testing others because of its location in a text or the particular construction of distractors. The editing teams, therefore, erred on the side of caution and tried techniques in their evaluation and revision after the two pre-tests.

4 Operation

This stage, together with its concurrent stage, Validation, brings us to the present day. The central activities are:

administration	- central v. local control, - management of resources, - expansion of the network,
marking	- monitoring objective answer sheets, - controlling subjective assessments.
score reporting	- average v. profile bands, - meaning of levels, - user requirements,
data storage	- needed for monitoring, - needed for validation, - needed for development.

4.1 A particular issue in the operation of ELTS is the control of eighty, soon to be about a hundred, test centres located in British Council offices which implement programmes particular to an individual country and as a consequence have decentralised management responsibility. Because of the demands of the 'chain' and the design thinking, the Service is a quite complex system of language tests yet it has to be capable of being controlled to ensure uniform administration and of being maintained not only with the provision of parallel versions but with revised, improved tests on the basis of the empirical evidence produced by the validation procedures. From the outset therefore it was planned as a testing service and not a series or battery of tests. To help achieve this a unit was set up in The British Council to monitor the implementation in 1980 and to liaise between the overseas centres and the various professional resource in Britain. Apart from improving the administration of the Service by revising guidance manuals, test report forms and providing training packages, the main developmental work is involved in setting up a schedule, and procedures, for further validating the present system and using that validation to answer questions such as does the Service need to provide further more delicate information - and so expand, or is much of this information redundant or 'unmeasurable' - and so contract.

5. Validation

This stage completes the cycle set out in the introduction to this paper but of course is, in many ways, a beginning and not an end. Because it is a test that affects the real life chances of so many people, ELTS has had to satisfy the standard demands and requirements made of such worldwide language tests. Because, however, it is required, by its many users, to provide information for such a variety of purposes, it has had to confront many of the major current issues in ELT: the specificity of ESP, language as process or product, definition of levels of ability, prediction of tuition time required to progress from level to level, etc. It is an operational system which provides initial, and needed, information on such issues and which is explicit and coherent enough to be capable of development. Validation studies will be the motor for such development and a number have been, and are being, set up. They all fall into one of the three areas set out below, the third area being the most important and the one expected to yield the most useful information.

5.1 Internal studies: analyses of test items (internal reliability and standards). These are run for each quarter's tests. The study includes: a study of the mechanical accuracy of scoring sub-tests and of calculating and recording bands; a re-marking study of the writing sub-test M2, designed to evaluate the reliability of marking and to refine any judgements about standards; a standard and routine item analysis of the sub-tests; a correlational analysis to see to what extent the various elements appear to be measuring the same or different traits: taken with estimates of reliability, this should contribute to an evaluation of the profile reporting; a study of standards as these were designed to produce the same band distribution in every element, to see if this aim has been successful.

5.2 Internal and external studies: analyses of sub-test specifications (content validity). The texts and items in the two General sub-tests and each Modular study skills sub-test to be re-described in terms of language skills, the relation to the 1979 skill descriptions and the original 1978 specifications. This to allow item analyses at 1 above to be related to the criterion specifications. The study will contribute to the control of the continuing process of specification, test realisation, validation and re-specification. It will also include a review of the face validity of the sub-tests and the revision of band descriptions.

5.3 External studies: validation of profiles (predictive validity). There will be a follow-up of all candidates in the first two years with information gathering by questionnaire to validate the band profiles, tuition hours, predictions and placement target levels. This will involve outside bodies (individual universities, SELMOUS, language schools) and departments in The British Council administering overseas students. The establishment of 'stereotype' profiles could then allow finer, more detailed studies with controlled observations. The systems used for gathering information will be designed to allow the variables to be processed on computer programmes at Test Development and Research Unit, Cambridge.

6. Conclusion

To describe the work of a number of people over a period of five years in a few pages is to run the risk of producing a narrative that sketches in only the general and does not come to grips with the particular. Those who are interested in the particular should refer to the publications of the Service which are available from The British Council. ELTS will produce, in the course of 1982, a Specialist Handbook which will be able to go into detail on issues about which this paper has provided only a brief summary. The Handbook will include the bibliography of reference material use in the production of the Service.

TEACHERS' ATTITUDES TOWARDS LANGUAGE TESTS AND TESTING

Douglas K. Stevenson and Ulrike Riewe (University of Essen)

ED217737

1 Introduction

As early as 1913, Cambridge University had an examination for English as a foreign language (Acheson 1977). And we are aware that in the 1920s, experiments with "new type modern language tests" were being published in book form in the U.S. (Jones 1980). But the birth of modern language testing is generally agreed to coincide with the publication of Lado's book in 1961. Before then in professional discussions about language testing, Lado, according to Ingram (1968: 147), "would have conferred with Lado."

Today - twenty years later - language testing has become one of the most popular and productive disciplines within applied linguistics. Many international conferences are now held each year. At least ten introductory-level textbooks on language testing are now available in English alone. A larger number of volumes containing conference papers exists. Others appear each year bearing imprints as varied as Singapore, Carbondale, Berne and Brussels. The growth of the field could also be measured by the Lange and Clifford (1980) bibliography which contains in its 350 pages more than 1,600 annotated references. Research into the cloze procedure only, that once "deceptively simple device," now rivals that accorded other staples of the academic publishing industry (James' *The Turn of the Screw* comes to mind). Certainly, those who would confer about language testing today quickly discover that there is no lack of willing correspondents. From the viewpoint of the rich mass of material now available, we as language testers might well conclude that during the past two decades, the field has gained a stronger voice, and has moved from being adolescent and dependent, to mature and influential.

Nonetheless, we owe it to ourselves to be more critical and objective when judging our own achievements, although we know that introspection is at least as dangerous as probing for long buried bombs, or opening best-forgotten love letters. We should, in examining our successes and failures, keep in mind the common goal, which has always been to improve the quality of language testing in language education contexts. Lado, it will be remembered, did not intend his volume primarily for psychometrists or other specialists. It was intended primarily for teachers of foreign languages and of English as a foreign language (1964:vii). His heirs for the most part have sought the same audience. Each has pointed out the benefits for both teacher and pupil that could derive from a better understanding of tests and testing.

FL013027

However, in an introductory textbook that has only recently appeared, written by Cohen (1980), one detects a difference in how this potential audience of teachers is addressed. For example, in the Foreword to the volume, Robert Cooper states that the book will be welcome "for language teachers who feel insecure about language testing," and he believes they "form a majority of the profession...." Cohen himself, in the Preface, refers to the many previous attempts to bring language testing and the teacher together. But he poses the question "to what extent do teachers actually consult such books on a daily basis?" Finally, there is the admission that is the first sentence of the first chapter, on the first page:"the word 'testing' sometimes frightens teachers because it suggests statistics and rigor."

Twenty years after Lado's Language Testing - and literally hundreds of articles, papers, studies, books, tests, research projects, seminars, conferences, and presentations later - it is appropriate to ask ourselves how successful we have been in reaching that primary audience. How many teachers of modern languages today still go into the schools without a thorough knowledge of tests and measurement? With so many good books on language testing at hand, why are there still so many horror stories about bad and inadequate testing in our schools? Why might the "majority of the profession" still feel insecure about testing, and why are many frightened by the word itself? Also, if we accept the well-known statement by Allen and Davies (1977) that "it would be curious to find an applied linguistics course in which testing and experimental methods were not taught," how do we account for the fact that, for example, during two recent semesters (SS 1980 and the WS 1980/81), among all the some 80 colleges and universities in both West Germany and Austria, only a total of three language testing courses were available? Jones has recently (1980) pointed out that training in research methodology is "essential" for advanced students in applied linguistics. How many of our students would agree with him? How many of their professors?

In this paper we are concerned with such questions. As language testers we are concerned because we suspect there is a discrepancy, even a gulf, between our own views and assumptions as to the worth of language testing to language teaching, and the views of those to whom we would primarily address our efforts. If such a gulf between views exists, but is not accurately surveyed, our best efforts might fall, and fail to reach the other side. We also feel that as more and more testers pursue their often esoteric interests, there is the danger that attention and effort will be drawn away from that common goal. It is important therefore that the question be asked and answers sought. Why, as we pass the twentieth anniversary of Lado's Language Testing, do the majority of language teachers still not receive adequate training in an area that we consider essential?

2 Expert opinion

Remarkably few language testing experts have concerned themselves with the question of why more good testing is modelled in textbooks than generally occurs in the classroom, or why more testers praise the benefits of knowing about testing than teachers are taught. Among those who do, however, three general arguments can be isolated (for a more complete treatment see Stevenson 1981).

The first reason, and surprisingly, perhaps the most commonly offered, is that despite the amount of material now available, little exists that is relevant to the practicing teacher's needs, or that is written at the level of the teacher-in-training. This argument is far easier to present than to substantiate. The fact that it is commonly encountered in whatever would be the latest successor to Lado's volume, and seldom countered, is interesting in itself. But as we have already noted, there are about ten introductory-level textbooks written in English alone. No one really knows how many exist internationally in all languages. Almost every one of them claims to be designed for the practicing classroom teacher, or teacher-to-be, while varying on which languages, educational levels, and systems receive the most emphasis. It is, as a matter of fact, difficult to find many books which are not of an introductory nature, and which are not directed at the teacher. Each recognizes that the practicing teacher has neither the facilities of an Educational Testing Service or a CITO, nor the authority of an examination board. Each assumes that very few teachers will have the statistical sophistication of a John B. Carroll, a Jum Nunnally, any Thorndike, or our own Ulrich Raatz. Certainly, none believe that teachers have nothing else to do than to spend their evening hours trying to write GRE-perfect test items; or by working out confirmatory factor analyses on the home computer; or if necessary, with a Number 2 pencil on a Big Chief Tablet on the kitchen table after the dinner dishes have been cleared.

Those who have advanced this type of argument also overlook the many language teaching journals which regularly carry easy-to-understand articles specifically written for the language teacher. There are also the many bibliographies and other introductory materials designed for teachers and teachers in training which are distributed by such organizations as the British Council, International TESOL, the Educational Testing Service, the Psychological Corporation and not least, the IUS. Once someone is introduced to language testing, s/he is very much aware of just how much material, written for the teacher, does exist. Whether or not the uninitiated, be it student or critic, is aware of this material is another question. Being available, of course, is not the same as being there, wherever the teachers are.

A second reason often advanced for teachers not being more familiar with language testing is equally susceptible to being specious, or self-serving. This argument claims that what is available is, pedagogically speaking, out-of-date. At its extreme, this argument pictures language testers as blindly following mechanistic, obviously ill-considered, methods and theories of the past. Testers stubbornly refuse to acknowledge present, revealed truths (which always vary considerably from prophet to prophet and from mountain-top to mountain-top). As has been pointed out elsewhere (Stevenson 1980), this type of argument achieves much of its effect by criticizing measures for not doing what they were not intended to do, and for doing precisely what they were intended to do. TOEFL and its close kin, for instance, have been attacked for not testing how well an examinee will perform in so-called real-life, language-use situations. Such a context would be, to give a familiar example, working as a head waiter at a tourist hotel in Spain. But TOEFL did not aspire to such purposes, finding it difficult enough (and perhaps impossible) to predict the ability to understand a lecture in English on medicine at the University of Michigan, to read advanced mathematics at MIT, or to write about multicultural education at New Mexico.

What does come across and rather clearly, is that a lack of attention to basic, well-established testing principles makes this second argument much easier to advance. A recent example, one deserving classic status, is that in a test of communicative ability, "reliability, while clearly important, will be subordinate to face validity" (Morrow 1979:151). Other also seemingly enduring misconceptions continually surface in many such arguments. A familiar one is that "objectivity" in testing has to do with content selection. Another still very much with us is that any reasonably intelligent observer can accurately determine what a test measures (often regardless of test purpose) simply by inspection. Still another is that no less a scholar than John B. Carroll condemned "discrete-point" tests for all purposes for all time to the paper shredder. Or, that "multiple-choice tests" are necessarily discrete-point tests. And then there is the what must be termed insidious misconception being widely propagated that certain testing approaches are somehow inherently reliable and possess some innate validity for all groups, purposes, administrators, and so on. Finally, for a field lagging far behind the latest trends, it is amazing how often its critics turn to those who have offended the most. Harris' suggestions for administration and scoring of oral interviews, for instance, are continually echoed in the very latest approaches. The grey-bearded FSI-type interview is being fêted as if it were a young revolutionary, wearing the green. And when Lado is read, some seem surprised to discover that he not only considered sociolinguistic situations and testing, but also predicted what the continuing problems would be, which are. Of course, it can be concluded, certain testers have emphasized certain approaches for specific purposes and testing contexts. Yet to claim that testing has nothing or very little to offer the teacher under that indistinct rubric, language as communication, is to rely more on rhetoric, than fact and familiarity with the literature.

The third reason sometimes offered for why teachers do not know more about testing is less commonly discussed, at least in public forums, for it involves certain sensitivities. Basically, this argument states that few teachers are required to receive specific training because many of the powers-that-be in teacher education feel that the water in the testing well is not really worth the trip. These individuals can be said to have a restricted view of language testing and its potential role in teaching. They do not necessarily say that what is available is not directed at teachers, or that it is out-of-date. Rather, language testing is simply not that central to the main concerns of the field. Any good teacher can also test well, and what the specialists amuse themselves with might be interesting, but is essentially peripheral.

What Johnson and Wong (1981) have termed a "historical antipathy" between teachers and testers also appears to be involved. We in testing tend to forget that modern measurement is a newcomer to language teaching and applied linguistics, in general. As H.H. Stern and Cummins recently acknowledged, it was not until the late 1950s that research, as opposed to expert opinion or simple authority, "began to affect in any truly significant way the policy decisions and the method debate in foreign language education" (1981:196). Modern measurement approaches are still new enough to language studies to be viewed with skepticism, resentment, or even hostility by those who were trained in traditional prose approaches. Furthermore, as Kuhn has shown, masters tend to teach their apprentices what they themselves once had to master, and continue to do so, until the pressure to change reaches a breaking point. For similar reasons, those who still view literature as the central component in foreign language teaching - and they still strongly influence or even control many language teaching systems - are not overly sympathetic to teaching their teachers about testing. Testers and their quantificational methods are sometimes viewed as the footsoldiers of the other of the two cultures, and as such, they are not welcome in the living rooms of the first.

We strongly suspect that of the three arguments outlined above, the last one best answers the question of why testing is not more familiar to more teachers. Language testing and measurement is frequently resented because it is a recent force in all language studies, and one that tends to challenge authority and change approaches. Relatedly, modern measurement is sometimes viewed as an attack on humanistic principles: testing is a field that would destroy or curtail creativity, and institutionalize or dissect, and therefore kill, insight. Yet all three arguments are based upon expert opinion and academic hunch. What is there that approaches the question of teachers' attitudes towards testing and goes beyond this level?

3 Available research

Unfortunately, there appears to be very little empirical research available, in spite of the importance of the problem, and against the expectation that testers, like professional gardeners, would do a lot of their digging in their own backyards. Three studies can be said to be typical of what we have been able to locate. Acheson (1977) surveyed teacher preparation programs for ESOL in American and British colleges and universities. Language testing was but one of several areas of interest, and the response rate was not optimal. Nonetheless, the survey did seem to indicate that despite popular preconceptions, little meaningful difference exists between the two traditions. Academic graffiti to the effect that the Yanks are all raving Chomskyites with biserial tendencies, or that the Brits are all amateurs who have no choice but to stay with the surface structure of language teaching have no basis in theory and practice. A somewhat surprising result is that in the U.S., so-called 'heartland' of educational and psychological measurement, only a fifth of the some 40 EFL-Master's programs which responded required coursework in tests and measurement. Of the only nine British programs of a comparable level, on the other hand, six required such work. The fact that some U.S. universities require that research skills be demonstrated at certain levels regardless of the area, and that the average North American student has a much larger and intimate experience with standardized testing at all academic levels, makes comparisons very difficult. Also, as Acheson pointed out (private correspondence), the British programs tended to rate tests and measurement fairly low although requiring it more often. Both traditions ranked the "integrative testing of communicative ability" higher than the "discrete-point testing of the four skills" and at virtually the same levels, whatever that means.

In 1974, a team led by Charles Stansfield (1976) sent a 125-item questionnaire to 3,500 foreign language department chairmen in the United States. Seven of the items dealt with textbook publishers' tests. Over 1,400 questionnaires were returned. One of the most surprising results is that two-thirds of the teachers claimed "never to use any tests they had not themselves prepared." Also somewhat surprising is that over 80% apparently did not like tests with multiple-choice items:

> Despite the prestigious position in American education occupied by objective tests, and their apparent advantages, the foreign language department chairmen surveyed rejected them overwhelmingly.

Apparently, according to Stansfield

> foreign language teachers were never really attracted by the idea of testing a single linguistic problem objectively as described by Lado and Valette, but they are common users of a more integrative approach. (113)

Stansfield, however, is clearly assuming that the teachers are familiar with Lado and Valette, in the first place, yet no evidence appears in the report to support this assumption.

A third study, interesting for different reasons, was carried out by Ronald Forrest (1975) who send questionnaires on objective tests to teachers in Ghana, Nigeria, and Ethiopia. The 86 replies show "that no matter how much the essential distinction between teaching and testing is stressed," teachers felt that objective tests were things to be taught and taught to. This survey, therefore, while not seeking attitudes towards language testing, revealed that popular misconceptions do affect judgments about certain testing approaches among teachers.

Past such surveys we have not found anything which tackles teachers attitudes towards testing directly and which also follows a good research design. On the other hand, Madsen, Jones, and Brown (1980) have recently reported some very interesting research which deals with attitudes toward test types. They demonstrate that students (of German) do possess definite, evaluative reactions to various kinds of tests, and that affective factors differentiate more distinctly among test types than do cognitive factors. Also important is their survey of the literature: the researchers were able to find little else that dealt with attitudes towards language tests (as opposed to testing), although similar research has a very strong tradition and impressive body of literature in the field of educational and psychological measurement.

We are more concerned, however, with teachers' and future teachers' attitudes toward testing, and understandably, the attitudes and views of those in our immediate educational tradition. Our students, unlike their transatlantic counterparts, do not come into the classroom already having taken hundreds of standardized tests in their previous educational careers. The various experiences and attitudes cloaked by such names and acronyms as the Nelson-Denny, Strong, ACT, SAT, WPPSI, WISC, WAIS, SCAT, CAT, MCAT, Millers, LSAT, GRE, or even the ubiquitous MMPI, cannot be presumed or held as inconsequential. In short, cultural expectancies are different. We don't feel that most or our students, asking a professor, "what will be on the final examination?" would find the answer, "all of the above/none of the above" meaningful, or amusing.

What then do our students, future language teachers, and our practicing teachers think about language testing? Past what we think they think, the educated hunch, we discovered that we simply do not know. And asking questions such as "Do you think this language testing course has been worthwhile?" usually tells the obvious. Those who do enroll in the few language testing courses available reward their own behaviour and, unless we assume students are hopelessly naive, that of those who taught the courses. At this point, and given this background information, it was necessary to decide if the problem was worth pursuing in even more detail. We could, as language testers, continue to write more materials and continue to hope that more students and teachers than previously would actually consult them. But twenty years after Lado's Language Testing, more of the same is less appealing. Importantly, each passing year sees a new generation of teachers go into the schools to support that present majority of teachers who do not have that essential testing knowledge. Only a relatively small number of practicing teachers can be reached through special seminars and other forms of continuing education, in spite of the best efforts of groups such as the AKS or IUS. To expect teachers once they have left the university to start reading books on testing, on their own, is also not the height of realism. We might remind ourselves of the books and studies we were going to read just as soon as we had time, during the summer, that has always just past! For such reasons, the problem did seem well-worth pursuing. If we are truly convinced that testing can do much to improve the quality of teaching, then once we have determined existing attitudes towards the subject, we should also be able to attract more students to the courses now available, bring more courses into more curricula, and better design continuing education programs and materials to fit teachers' views and whatever preconceptions might exist.

4 Informal interviews

As a first step, informal interviews with future teachers of

English seemed appropriate to determine what the broad boundaries were for later and more specific attitudinal measures. They would also indicate whether or not expert opinion had any firm basis in fact. Moreover, the interviews would have to be non-overt by both trait and method. That is, those situations would have to be used in which a conversational question about language testing, as a subject, would not be inappropriate. The one who would essentially take advantage of the situation would therefore have to be a student, as well. We wanted to see if the question itself existed. To present it overtly and directly by questionnaire or formal, structured interview, or by someone identified as either tester or teacher/professor could easily bias the results. Therefore, a student often present during such conversational situations, yet familiar with both the background to the research and attitudinal measurement, was cleverly "disguised" as a normal student. Needless to say, this technique does not allow generalizations from the results, the claim that they are representative. Nor, it should be noted, do we claim they are such. Rather, the 16 "near normal" conversations have proven useful in determining what must be included in more closely controlled, and more precise efforts.

Two general groupings of responses emerged: further classification, given the methodology, is not appropriate. First, and supporting the appropriateness of a non-overt approach, many of the students seemed to have no opinion. Or perhaps more accurately, the question was a non-question, answered in the way we might respond to one we'd not previously even considered. A typical response in these conversations was "Language Testing? Als Lehrer brauche ich das doch nicht, oder?" (Language Testing? It's not important for me as a future teacher, is it?) We found such responses rather depressing, for one would have expected that at sometime, at some stage of their preparation in some course, attention would have been given to the importance in the classroom of good language assessment. Given the amount of testing activity at this university (and a strong departmental support for this activity), this response was simply not expected. Nor did expert opinion, already referred to, predict it. Perhaps this is one indication that our basic assumptions as testers are hardly as well known, let alone shared, as we might wish to believe.

The second general response was anticipated, although not sought. Here "tests" are those "M-C" devices which are fundamentally oppressive, even "fascist," and designed to unfairly categorize or restrict pupils and students from doing what they could do. Such devices, it would appear, are so obviously ill-construed and even simplistic that it is quite obvious that they do not, and cannot, test anything worth knowing. Such responses echo the familiar exposés in the popular media. They are infamous, at least to language testers, for condemning bad testing for the wrong reasons, and for employing ridicule where knowledge is lacking. It is in fact noteworthy that so many self-assured opinions of this type exist among so many who have no real knowledge of testing. There are also obvious correspondences with various political positions. "Tests", for example, are seen as agents of those who would maintain class lines and further the already high level of stress that comes from a competitive society. This is a complex area, to say the least, but the fact that the responses are essentially negative indicates that those chapters in textbooks which outline the benefits for teaching - not to mention the reforming and revolutionary tendencies of testing - are not known. If they are, they have evidently been ineffective. At least for this small group, language testing is way down the list of appealing subjects. It would not simply be a question of presenting a strong case for language testing, as it would be with the first, "no-opinion" group. Rather, it would be like trying to find a defense attorney after the jury has already found the defendant guilty as charged.

At present, we are going ahead with the design for a large-scale study. In this study a variety of attitudinal measures, complemented by selected, structured interviews, will be used to better determine attitudes towards language testing among future teachers, practicing teachers, and representative senior professors who have a voice in determining the curricula for language teacher preparation programs. Through the IUS membership, cooperation has already been offered by five universities in Germany, one in Belgium, two in the United Kingdom, one in Sweden, and one in the United States. Further such support is expected, and most welcome. A so-called "do not turn back" format appears necessary, that is, certain item-sets will be answered and collected before others are distributed. Preference scales for various academic subjects, testing statements versus political statements matching items, and the semantic differential technique should give us more information about attitudes than has been previously available. We also hope to further develop the "normal student" interview to allow formal data and data interpretation. It is our belief that opinions coming from "real" questions such as, "Did you get one of those questionnaires? Did you fill it in? What did they mean by that?" when properly tabulated, can give a different view of attitudes than will the overt questionnaire results.

However, once such information is available, we will be in a better position to design language testing courses and materials which acknowledge and deal with attitudes towards testing.

Bibliography

Acheson, P. 1977. English for Speakers of Other Languages: a survey of teacher preparation programs in American and British colleges and universities. In J.F. Fanselow and R.L. Light (eds). Bilingual, ESOL and Foreign Language Teacher Preparation: Models, Practices, Issues. Washington, DC: TESOL.

Cohen, A.D. 1980. Testing Language Ability in the Classroom. Rowley, MA: Newbury House.

Forrest, R. 1975. Objective examinations and the teaching of English. English Language Teaching Journal 29/3, 240-246.

Ingram, E. 1968. Testing in the context of a language learning experiment. In J.A. Upshur and J. Fata (eds). Problems in Foreign Language Testing. Language Learning Special Issue Number 3. Ann Arbor, MI: Research Club in Language Learning.

Johnson, F.C. and C. Wong. 1981. The interdependence of teaching, testing, and instructional materials. In J.A.S. Read (ed). Directions in Language Testing. Anthology Series 9. Singapore: Singapore University Press.

Jones, R.L. 1980. Assessing second language proficiency: where are we now and where are we going? Paper presented at the Southern Illinois Language Testing Conference, Carbondale.

Kuhn, T.S. 1962. The Structure of Scientific Revolutions. International Encyclopedia of Unified Science 2/2. Chicago University of Chicago Press.

Lado, R. 1961, 1964. Language Testing: The Construction and Use of Foreign Language Tests. London: Longmans, Green and Co. New York: McGraw-Hill.

Lange, D.L. and R.T. Clifford (eds). 1980. Testing in Foreign Languages, ESL,, and Bilingual Education, 1966-1979: A Select, Annotated ERIC Bibliography. Language in Education: Theory and Practice Series 24. Arlington, VA: Center for Applied Linguistics.

Madsen, H.S., R.L. Jones, and B.B. Brown. 1980. Evaluating affective variables in second-language testing. Paper presented at the Fourteenth Annual TESOL Convention, San Francisco.

Morrow, K. 1979. Communicative language testing: revolution or evolution? In C.J. Brumfit and K. Johnson (eds). The Communicative Approach to Language Teaching. London: Oxford University Press.

Stansfield, C. 1976. Teachers' attitudes toward publishers' tests. In D.P. Benseler (ed). Second Language Teaching 76. Proceedings, Pacific Northwest Council on Foreign Languages. Corvallis, OR: PNCFL.

Stern, H.H. and J. Cummins. 1981. Language teaching/language research: a Canadian perspective on status and directions. In J.K. Phillips (ed). Action for the '80s: A Political, Professional, and Public Program for Foreign Language Education. Skokie, IL: National Textbook Company.

Stevenson, D.K 1980. The sociolinguistics of language testing: a tester's perspective. Paper presented at the Internationaal Colloquiun Taaltoetsen, Brussels.

- 1981. Language testing and academic accountability: on redefining the role of language testing in language teaching. IRAL 19/1 15-30.

ED217738

FL013028

A STUDY OF THE PREDICTIVE VALIDITY OF THE MICHIGAN BATTERY, WITH SPECIAL REFERENCE TO THE COMPOSITION COMPONENT

Dianne Wall (University of Lancaster)

1 Introduction

This paper describes an attempt to follow up the performance of foreign graduate students in the College of Engineering, University of Michigan, in Autumn 1979, and discusses whether the scores they obtained on the Michigan Battery, with special reference to the composition test, predicted academic success or lack of success in their first-term subjects. It also presents an analysis of the writing demands the students were faced with in their Engineering classes and a comparison of these demands with the writing task they had to do for the composition test. The paper will conclude with implications for future test design.

2 The Michigan Battery

The Michigan Battery is an English language proficiency test which is produced and scored at the English Language Institute, University of Michigan, and is administered to approximately 8 000 students a year. According to the Manual for the Michigan Test of English Language Proficiency, one of the three components of the Battery, scores obtained on the Battery are to be used

> in estimating whether a student whose native language is not English is able to pursue academic study in an English language college or university, and how much study he might be able to undertake at his present level of proficiency in English. (English Language Institute 1977:1)

Admissions offices in colleges and universities in the United States and Canada regularly request that student applications be accompanied by scores from either the Michigan Battery or the Test of English as a Foreign Language (TOEFL), to demonstrate that students have a sufficiently high level of English to manage the coursework that will be expected of them in their academic careers.

The Michigan Battery is composed of three parts: the aforementioned Michigan Test of English Language Proficiency (MTELP), which is itself broken down into grammar, vocabulary, and reading comprehension subsections; a test of aural comprehension (if local conditions do not favour this taped test, an interview is substituted); and a written composition. The scores for each part, as well as the total Battery score, are sent to the institutions specified by each testee.

Also sent with the scores are guidelines for their interpretation, officially called "Proficiency Recommendations". According to these guidelines a graduate student in Engineering who scores a Michigan Battery total of 96-100 should be able to "compete with native speakers on equal or nearly equal terms" (ibid: 12), whereas students with scores of 84 or lower should have some restrictions placed on the number of courses they can enrol for and should be required to take supplementary English classes.

The evidence supporting these recommendations is scant. Despite the fact that the Michigan Battery is widely used for selection and placement decisions, the authors of the test have done little follow-up work on the students who are placed by their instrument. According to the MTELP Manual, the only follow-up studies that have been carried out took place in 1956-57 and 1967. Unfortunately, results of the first study have little relation to the current test's predictive validity, since the Michigan Battery at that time was composed of different tests than the current Battery and the Manual gives no information on how those tests might have compared with the ones used now. The 1967 study reports the correlation between Michigan Battery scores and grade point averages of graduate students in the humanities and social sciences; however, no correlation is reported for students in Engineering.

The authors excuse their lack of follow-up by stating that

> there have been too few cases in which recommendations (their own recommendations for admissions cut-offs) have been waived to allow a meaningful determination of the efficiency of the test as a predictor of academic success. (ibid: 14)

It is true that any follow-up study of students who have been placed by the Michigan Battery will not be able to consider the performance of students who have been refused admission on the basis of their proficiency scores; however, this does not mean that work cannot be done to investigate whether students with "attested" proficiency have been able to cope with the demands placed on them in their study environments.

3 The composition component

When taking the composition test of the Michigan Battery the student is given 30 minutes to read and analyse a topic (if two topics are presented he must also select the one he prefers to write about), organise and write his response (in 1979 he was requested to write 250 words), and revise and correct his work. The topics are non-subject specific, and the task itself can be considered "free writing" in that there is only a minimum of input provided for the writer. The papers are marked by two readers, who read them quickly to form an overall impression and then assign them marks on a 10-point scale. The composition mark reported to admissions officers is usually the mean of these two marks.

According to the MTELP Manual, the purpose of the composition is

> to show whether a student is capable of handling English well enough to indicate his knowledge of subject matter in a course examination type situation. (*ibid.*:1)

The Manual explicitly states that this "is not a comprehensive test of written production of English", but does not indicate what the criteria are for judging the student's written output. Indeed, no rationale has been published to explain why readers mark the way they do; nor is there any published account of the factors considered by them to be most important to academic writing.

With regard to the predictive validity of the composition, no studies have been published.

4 Description of research

The lack of follow-up studies for the Michigan Battery as a whole and for the composition component in particular has serious implications, not only because the admissions and placement decisions made on the basis of the Michigan scores can prove time-consuming and expensive for the students involved, but also because the question of whether what is being assessed in the tests is essential to academic success is being ignored. It was therefore decided to do a follow-up study of overseas graduate students in Engineering, to see if their Michigan Battery scores predicted anything useful about their performance in the classroom in their first term of studies.

Graduate students were studied because university records for 1978-79 showed that 12% of the graduate student population had come from overseas, as opposed to only 4% of the undergraduate population. Students from the College of Engineering were selected because this college had the largest enrolment of overseas graduate students: 441 out of 1 096, or 43%. The investment being made by the students themselves, their sponsors, and the Engineering faculty emphasised the need to predict as accurately as possible whether any student would achieve academic success. This screening was traditionally done by English proficiency examinations, but if it were shown that these were not effective means then resources could be directed to finding other instruments for that purpose.

The subjects who were selected for the study were students in Civil, Mechanical and Electrical Engineering, and all of them were in the first term of a year-long MSc programme. Though their specialties differed, they shared a number of minimum requirements for a degree, one of which was to maintain a B- average throughout their year of study. This cumulative grade point average (GPA) was therefore selected as the criterion of academic success against which Michigan Battery scores would be correlated. All of the subjects had sat for the complete Michigan Battery, and their total scores ranged from 73 to 96.

4.1 Correlation study

A Spearman rho rank correlation coefficient was calculated for first-term GPA and Michigan Battery scores for the 13 students. The results can be seen in Table 1 below.

Table 1

Michigan Battery scores and grade point averages

	GPA	Total	Comp	MTELP	AC
GPA		.31	.37	.31	.22
Total	.31		*.72	*.94	*.85
Comp	.37	*.72		.53	.38
MTELP	.31	*.94	.53		*.82
AC	.22	*.85	.38	*.82	

N = 13 * Significant at the .01 level

No significant relationship was found between the GPA and the total Michigan Battery score, or between the GPA and any part score. The only significant relationships existed between the part scores and the total, and between the MTELP and the aural comprehension test, which suggests that there may be some overlap between them. Though the composition itself was not found to be a good predictor of academic success, it seemed important to investigate the possibility that the predictive validity of the total Battery was enhanced by the inclusion of an integrative component in an otherwise discrete-point test. It was not possible though to show this was the case, since neither the total including the composition nor the total excluding it correlated significantly with the GPA.

4.2 Discussion of writing demands

Investigations were also carried out into the writing demands placed on the subjects in their Engineering classes, and the criteria used by their tutors to evaluate their written work. It was thought that if the composition test was not predicting how students would perform in their written work, then these studies might suggest reasons why not. In addition, an analysis of classroom writing might suggest new formats for the composition component, raising its face validity.

Interviews were therefore held with the subjects' Engineering tutors. Each tutor was asked to answer a number of questions about how much writing the student was required to do, the form that writing was to take, and what the tutor was looking for when he evaluated the students' written work. He was also requested to provide a sample of "typical work" from his classes, which could take the form of a homework assignment, an examination, a lab report or project. The word "typical" was deliberately left vague, since requesting a sample "with writing on it" might have produced a distorted picture of classes for which little writing was done. Of the 40 tutors interviewd, 30 complied with a piece of written work.

The interviews revealed that students in their first term were required to turn in much written work, and that it was often the written work rather than oral performance that determined their final marks for a course. However, a review of the samples showed that much of the writing students produced for their Engineering classes had a different function from that produced for the composition test. For example, in a typical Engineering assignment writing was used to guide the reader through pages of numbers and figures. These calculations and diagrams carried the main burden of the message that the student wished to project and what is traditionally called "writing" provided only reference points in an otherwise non-verbal process. Not only the function but the nature of the writing was different. Within the reference points themselves it was not necessary to produce "correct" writing, except in terms of its truth value. Grammar did not seem crucial given that language was often manifested in small chunks, and whatever cohesion or coherence there was in a paper might depend more on the mathematical operations being illustrated than on the writing that accompanied them.

Of course some assignments did require connected prose in the form of definitions or explanations, ranging from several sentences to several paragraphs, but tutors generally insisted that they were looking for specific elements in this writing and it did not have to be formally correct to be communicatively effective.

4.3 Analysis of underlying requirements

Since the composition and the coursework writing differed so much on the surface level, it was decided to approach a comparison from another angle - to analyse both types of work according to the underlying requirements involved in producing them. That is, it was intended to see if the processes a student went through while writing a test composition and writing for Engineering had elements in common, which might justify the testing of composition the way it had traditionally been done. If the analysis showed that the underlying requirements were completely different, then perhaps the elements present in the Engineering assignments could be incorporated into a new composition task.

The term "underlying requirements" had to be limited to behaviours that could safely be inferred from an examination of text, such as "decide on something to write about", rather than hidden psychological processes such as "infer" or "associate". It was felt that proposing the latter would only be speculative, and the idea of incorporating any "findings" into a test design - that is, trying to ensure that subjects use the behaviour in question in the writing of a composition test - seemed both dishonest (How could it be said that the essential behaviours had been pinpointed?) and impractical (How could one ensure that the desired behaviours would be employed?)

4.3.1 The Engineering tasks

In the analysis of any written work the writer's medium, audience, purpose and situation must be taken into consideration. In their coursework the Engineering students could communicate through three media: numbers, visual aids such as diagrams, and words. They were aware of who their audience was: an expert, whose function it was to test their capabilities. Their purpose was to prove to that expert that they understood what had been covered in class and the texts, and could transfer this to new problems. Their situation, how they were to approach their work, varied according to the task, but several factors remained constant: they would be able to rely on previous input from lectures and reading, they would receive explicit instructions, the variables they would have to deal with would be included in the rubric of the task or would be retrievable from sourcebooks, and they would act on their data in a conventional way and would try to present correct or plausible answers according to a standard format.

These and other observations are discussed in detail below, and illustrations are given from the sample work analysed.

a. Understand previous input. The subjects were only asked to tackle problems for which they had been prepared by previous reading, lectures, or discussions with tutors and other students. When they prepared homework assignments they were allowed to refer to notes, texts and reference books, and when they were in examinations they could rely on their memories. They were not required to deal with theories, methods or terminology that they were not already familiar with or which they could not find in source materials.

> Example: "Describe the responsibility of the project manager on a project which also has a project superintendent." (Student could rely on information from readings.)

b. Understand what was expected. The Engineering assignments were explicit in their requirements: calculate, find, draw, plot, etc., and the tasks were ones that the students were technically capable of doing. Each student was expected to prove that he could carry out the necessary operations and was not encouraged to elaborate on them or do any uncalled-for work.

Example: "Determine the operating voltage, operating current, and required load resistance to obtain maximum power from the cell into the external load. Give the d.c. power obtained."

c. Interpret numerical, visual and written verbal input. In addition to the data alluded to in Point 1 above, the student received more input in the actual task instructions, present in the form of descriptions, diagrams, charts, etc. All the relevant details were given to enable him to complete his task; he was not expected to create variables for himself or use his imagination to provide a context for the problem.

Example: "A Mitatt diode has a generation region width x_a equal to 500A, with a constant electric field E_a of 9×10^5 v/cm in the generation region ..."

d. State models and assumptions. The student was required to specify how he planned to apply theory to reality, though this did not have to be in sentences. The presence of an equation on the page was evidence that it would be used; the student had merely to label the theory and method he was employing to make the task of following less difficult for his reader.

Example: "Assuming that the neutral axis remains at the midthickness of the bent sheet, then

$\frac{R}{t} = \frac{50}{Ar} - 1$, where R = radius of bending
t = thickness of the sheet
Ar = area reduction of materials."
(student response)

e. Manipulate numbers to arrive at correct or plausible answers. Several of the tutors interviewed made statements like "Our language is algebra" and "Equations are words", and the samples showed that the two media were comparable in many respects. As he would with words, the student could use numbers to capture his thoughts symbolically, and these numbers could be interpreted by a reader removed in time and space. However, manipulating numbers and manipulating words proved different in one respect: that with numbers there was only a right and a wrong. Of course it was possible to find some answers that were more "elegant" than others, but even the concept of "elegance" depended on whether the answer given matched the one considered most appropriate by the tutor.

Example: "Using the information given above, calculate a) the hourly ownership cost and b) the hourly ownership and operating cost."

f. Capture a process as product. The product that was required on many assignments was the process that the students had had to pass through in order to solve a problem. It was necessary to record the process to prove that one had not arrived at the acceptable answer by chance, and to refer back to in case the solution was not the right one. This process involved the manipulation of a limited set of variables and was therefore capturable on paper. The writer was not expected to make the process interesting, but only to demonstrate that he had gone through it without being sidetracked by false assumptions, inaccurate operations, etc. He was required to empty his mind onto the paper; that is, to show all the steps, in the right sequence, with no sudden jumps between them. If he could not deal with a certain operation, he could not continue his written work, since every operation and result fed into what followed.

In this attempt to recreate process on paper it was necessary to coordinate numbers, figures and words, and here the subordinate role of writing became evident: the language that was entered on student papers helped the reader find his way in the process but it could not take over if the calculations failed.

g. Illustrate. The student needed to support his work with instantly interpretable visual displays: diagrams, charts, tables, circuits, plans or flow charts, and had to label them appropriately.

h. Follow conventions. All the information had to be presented according to conventions laid down by the tutor or the standard texts. The student was not expected to be original in his presentation, but rather, was to present his data in a concise, complete, straightforward way.

i. Evaluate. There was sometimes a need to discuss results: interpret them, justify them, explain how they varied from textbook models. This was one of the few things that had to be done in writing, though tutors were generally in agreement that if an explanation could be followed they would not be disturbed by language problems within it. Tutors also pointed out that making an explanation "followable" was not necessarily difficult, since the tutor, the person who had to understand, was usually already familiar with the situation the student was writing about and with the problems he was having. Even if the student's language was not as precise as it might be his arguments were often understandable since the tutor could supply many of the details for himself, from his own experience with the subject matter.

j. Document work. The student was expected to refer to source material and acknowledge it.

Example: "The report should conclude with a numbered list of cited references."

k. Be accurate. The student had to not only give the correct answer but make sure that his calculations were correct throughout the process.

l. Be neat. Some tutors deducted points for messiness and gave extra credit for "professional presentation". The students not only had to record their thoughts as they were thinking them, but ensure that they could be deciphered at every stage.

In short, when Engineering tutors marked written work, they paid attention to the following: how well the student understood the concepts and the relationships between them; the followability of the process - whether the steps were in the right order and the work was accurate; whether the student arrived at the right answer; and whether the work was presented in a professional (i.e. tidy and well-organised) way.

4.3.2 The composition task

Of the subjects selected for this study, all but two were given the Michigan Battery the same day and had the same choice of topics:

> "The best defence against war is that every country have a large army." Do you agree or disagree? Why?

or If you could live at any other time in the history of the world, when and where would you live? Why?

One subject who was tested a week later was given this choice:

> How should students be evaluated? Should they be graded on achievement or effort? Should they be given grades or marks at all? Discuss.

or What is the funniest thing that ever happened to you?

The final subject, who was tested several months before the others, had no choice of topics and had to write on the evaluation topic given above.

The mark a student receives for his composition is not supposed to be influenced by his choice of topic, though it is evident that the above tasks all require different talents of a writer. The topics concerning war and evaluation both require that the subject argue for or against a proposition, but the latter gives more guidelines for approaching the task. The topic about living in another time requires no argumentation, but only the description of preferences, while the topic about funniness demands not only a simple narration in the past but that the writer remember something that will seem funny to someone else. In 1979 there were approximately two dozen topics which were presented in different combinations (when a choice was given to testees). These involved several modes of writing - argumentative, persuasive, descriptive, narrative - and required that the student produce his own input as well as organise his thoughts and put them down on paper in a grammatical way.

It is not possible to specify what an ideal response would look like for these very varied tasks; however, a list of underlying requirements common to all topics can be given to compare with that drawn up for the Engineering samples.

In contrast to Engineering assignments, the composition allowed only one medium of expression: words. The audience was rarely specified so that subject had to decide who he was writing for and what concessions to make to that particular person. The purpose of the writing was also left unclear. The writer could guess that he was supposed to demonstrate "good writing", but what did "good writing" involve? Good grammar? A good idea? A good sense of humour? Lastly, the subject's situation was quite different from the one he would face in Engineering, as discussed below.

In order to deal successfully with a topic (receive a high mark from two holistic readers), a subject would have to:

a. Rely on his knowledge of the world and other intangible things. A subject receives very little input from the topic itself and must spend much of his 30 minutes planning something to say. It may be that the student who has an interesting past, more imagination or is "inspired" may be favoured by this lack of standard input. It should also be noted that some writers may not be able to write about some topics even in their native language - they may never have had cause to think about them, or may feel threatened emotionally or politically by the demand to disclose their feelings about the question.

b. Decide on the argument, incident or opinion to be presented. This may involve weighing factors like "What is the easiest thing to do?" or "What does the reader want to hear?", both of which take time away from the choice of words or actual mechanics of writing. Deciding on the answers is of course complicated by not knowing who the reader is, or what the writer himself is supposed to be demonstrating. Students who are ambitious or want to communicate complex ideas may run the risk of over-extending themselves linguistically, without knowing if the quality of the idea or the quality of the sentence structure will determine the mark.

c. Decide where to begin and how to continue. Since the situation is not set up for the writer in the instructions, he must decide which of his ideas are essential and which irrelevant, as well as how to sequence them to provide for maximum impact on an unknown reader. The number of variables involved in writing and the fact that there is no correct solution make the "capture a process as product" idea untenable. Whereas the Engineering assignment allows the writer the "freedom" of following his own processes, in the "free writing" task he may find the need to make so many choices disabling. That is, he may have to spend so much time in the pre-writing stage that he will not produce enough language to demonstrate his writing competence.

d. Fill in gaps in the reader's background knowledge. Since the writer cannot assume that he and his reader have background knowledge in common, he should be careful to anticipate his reader's questions and fill in the details that are prerequisites for understanding. Unfortunately, precisely because he lacks knowledge of the reader's knowledge, he may not know what concepts and situations need clarifying.

e. Be correct. The criteria that were suggested to readers-in-training prior to mid-1979 stressed fluency (undefined), complexity, broad vocabulary and lack of discrete errors as characteristics of high-level compositions (English Language Institute 1979:3). Factors generally taken into account when marking coursework writing were not explicitly mentioned, such as organisation, relevance to topic, truthfulness or appropriacy. These factors may indeed be taken into consideration at present, yet as mentioned previously no account has been published of what the authors consider "good writing" to be or why they test writing the way they do.

f. Be fast. The writer has only 30 minutes in which to work, and he must not only organise his thoughts but write them down on paper and correct them. It is important to note that none of the topics are piloted with either non-native or native speakers, "good" writers or "bad"; therefore, it is not known whether the task required can be successfully completed by even proficient L_1 writers in the time allowed.

5 Summary of differences

A summary of the differences between the two tasks can be seen in Table 2.

It can be seen that there are many differences between what is required for the composition test and what is required for written work in Engineering. The main difference seems to be that in the Engineering tasks there is much prior input and the task itself is explicitly outlined, whereas in the composition the writer has only a suggestion to respond to and must not only create the content of his writing but a context, audience and purpose as well. The criteria for marking would also seem different, although those currently used by composition markers would have to be articulated before this debate could be continued. In the past the criteria have been somewhat vague. For example, in order to obtain the second highest mark a student needed to "be as good as an 18-year-old native-speaker college student"; however, it should be recalled that no pretesting is done to see just how good that is.

6 Conclusions

Two analyses were carried out on data received for overseas Engineering postgraduates. The first was a correlation study between the Michigan Battery total and part scores and the students' first-term GPA, in which no significant relationship between tests and the criterion for academic success could be found. Several factors could prevent a high correlation from occuring, such as the aforementioned need to use a truncated sample, grade inflation, the number of technical subjects taken by each individual, the amount of English tuition and its effectiveness, and various socio-affective factors. The authors themselves state that "it should be recognised ... that language performance is only one of the variables affecting student performance." (English Language Institute 1977:14);

Table 2

Summary of the differences between the written work produced for Engineering courses and the composition component.

Engineering Courses	Composition Component
Media: numbers, visual aids, words	Medium: words
Audience: known expert	Audience: must be invented
Purpose: to prove one can solve a problem adequately, according to established criteria	Purpose: to prove one can "write well", according to unknown criteria
Situation: much prior input, directly concerned with problem to be solved	Situation: reliance on knowledge of the world, imagination, inspiration
explicit instructions, limited task	suggestions for "free writing" task
numerical, visual and written input within the task rubric	need to create a situation and variables
need to state models, assumptions	need to decide which argument, incident or opinion to present
manipulate numbers to get right answer	no right answer
capture process as product, step-by-step narration	decide where to begin and how to continue
illustrate with visual aids	fill in gaps in reader's background knowledge using words
follow conventions specified by teachers or texts	follow conventions, if known
be numerically precise throughout	use good grammar
be neat	be fast

however, it does not seem fair to state that on the basis of present evidence that the English proficiency the Battery claims to measure, constitutes even one of the variables affecting success in a subject like Engineering.

With regard to the second analysis, a comparison of the underlying requirements involved in writing for the composition test and writing for the majority of Engineering assignments examined, it was found that they differed considerably, mainly in terms of the amount of input that the subject was expected to create for himself in order to write the composition. As the whole task had to be completed in only 30 minutes, it seemed to be biased in favour of students who could organise their thoughts more quickly, a questionable goal for a test that claims to be assessing language proficiency.

It was concluded that the composition may be testing more than it needs to, including factors such as originality and personality, rather than just how effectively the student can "indicate his knowledge of subject matter in a course examination-type situation." In The philosophy of composition (1977:187) Hirsch states that "far greater variations in quality of ideas or aims will occur than in quality of presentations", unless extrinsic factors required in a composition are controlled for. This can be done by providing a context for the writer to fit his writing into. The marking that would follow might then attempt to measure these factors, as suggested by Lado (Lado 1961:296):

a. the inclusion of essential points of information
b. the organisation and sequencing of such information
c. the use of formal signals to guide the reader through the text.

The composition assignment would not have to replicate course assignments in content, since it is the student's ability to present information rather than the content itself that needs to be assessed. The test need not even be subject specific, but merely provide the student with some sense of audience and purpose, and raw material to organise - saving him time and allowing his readers to focus on his communicative ability, rather than how well he "thinks on his feet". A new format might be suggested by tests of writing such as the Joint Matriculation Board Overseas English Examination or the writing assessment of the English Composition Board, University of Michigan.

The criteria for judging such writing should also be discussed and open for discussion. It is necessary for test-makers to come to grips with the question of what good writing consists of, and consider what the fairest way of presenting tasks and marking them might be. The problems that arise here are numerous, but hopefully the message is clear: the same care should be taken in designing the subjective elements of a language proficiency battery as in designing the objective sections.

Though the preceding discussion has concerned one particular proficiency battery, the remarks about composition testing apply to other testing programmes in Britain as well as in the United States, and to the testing of writing in the L_1 and L_2 classroom. The larger issue, that of the function and ideal nature of language proficiency tests, has been the subject of much debate and could not be expanded upon in this limited article.

Bibliography

English Language Institute. 1977. Manual for the Michigan Test of English Language Proficiency. Ann Arbor: English Language Institute.

- 1979. "Some characteristics of different level compositions". Ann Arbor: mimeo.

Hirsch, E.D. Jr. 1977. The philosophy of composition. Chicago: University of Chicago Press.

Lado, R. 1961. Language testing. London: Longman Group Ltd.